Derivation of the EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models

Elizabeth Eli Holmes
Northwest Fisheries Science Center
Mathematical Biology Program
NOAA Fisheries
2725 Montlake Blvd E., Seattle, WA 98112
eli.holmes@noaa.gov
http://faculty.washington.edu/eeholmes/

April 15, 2010

Contents

1 Overview

EM algorithms extend likelihood estimation to cases with hidden states, such as when observations are corrupted and the true population size is unobserved. EM algorithms are widely used in engineering and computer science applications. The reader is referred to (McLachlan and Krishnan, 2008) for general background on EM algorithms and to (Harvey, 1989) for a discussion of EM algorithms for time-series data. Coding an EM algorithm is not as involved as the following 30+ pages might suggest. In most texts, the majority of the steps shown in this technical report would be subsumed under the line "it follows easily that...". However, if one has never derived an EM algorithm or update equations for multivariate normal models, the steps might not be so obvious. This technical report covers each step in detail so that those who wish to derive an EM algorithm for extensions to the MARSS model can see the exact steps and logic required.

The EM algorithm that is presented in textbooks is for the unconstrained MARSS model where all parameters elements are estimated. In the Mathematical Biology group at NWFSC, we work mainly with constrained MARSS models where there are fixed and shared values throughout the parameter matrices. An example of a shared value would be a shared growth rate term (u) across all state processes (the random walks) in a MARSS model. In this report, I review the derivation of the unconstrained EM algorithm and then show the derivation of the constrained MARSS update equations.

Our linear MARSS model is

$$\boldsymbol{x}_t = \mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{u} + \mathbf{w}_t, \text{ where } \mathbf{w}_t \sim \text{MVN}(0, \mathbf{Q}) \tag{1a}$$

$$\boldsymbol{y}_t = \mathbf{Z}\boldsymbol{x}_t + \mathbf{a} + \mathbf{v}_t, \text{ where } \mathbf{v}_t \sim \text{MVN}(0, \mathbf{R}) \tag{1b}$$

$$\boldsymbol{x}_1 \sim \text{MVN}(\pi, \mathbf{V}_1) \tag{1c}$$

Our derivation of the EM algorithm for the unconstrained[1] MARSS model is based on the derivation by Ghahramani et al.(Ghahramani and Hinton, 1996; Roweis and Ghahramani, 1999). This EM algorithm was originally derived by Shumway and Stoffer (1982), but our derivation follows Ghahramani et al's slightly different development[2]. Here, this derivation is extended to the case of a constrained MARSS model where there may be fixed and shared elements in the parameter matrices. The algorithm consists of an expectation step ("E step"), which computes the expected values of the hidden states using the Kalman filter/smoother, combined with a maximization step ("M step"), which computes the maximum-likelihood estimates of the parameters given the data and the expected values of the hidden states.

[1] "unconstrained" means that each element in the parameter matrix is estimated and no elements are fixed or shared.

[2] One difference is the treatment of the initial condition. The initial condition is $\boldsymbol{x}_1$ in our derivation not $\boldsymbol{x}_0$. The result is that our update equations are slightly different than Shumway and Stoffer's; although both lead to the same maximum-likelihood parameter estimates.

1.1 The log-likelihood function

Before describing the algorithm, we need to specify the joint log-likelihood of the data and hidden states for this model[3]

$$\begin{aligned}\log \mathbf{L}(\boldsymbol{y}_1^T, \boldsymbol{x}_1^T|\Theta) = &-\sum_1^T \frac{1}{2}[\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a}]^\top \mathbf{R}^{-1}[\boldsymbol{y}_t - \mathbf{Z}\boldsymbol{x}_t - \mathbf{a}] - \frac{T}{2}\log|\mathbf{R}| \\ &- \sum_2^T \frac{1}{2}[\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u}]^\top \mathbf{Q}^{-1}[\boldsymbol{x}_t - \mathbf{B}\boldsymbol{x}_{t-1} - \mathbf{u}] - \frac{T-1}{2}\log|\mathbf{Q}| \\ &- \frac{1}{2}[\boldsymbol{x}_1 - \boldsymbol{\xi}]^\top V_1^{-1}[\boldsymbol{x}_1 - \boldsymbol{\xi}] - \frac{1}{2}\log|V_1| - \frac{n}{2}\log 2\pi\end{aligned} \tag{2}$$

$\boldsymbol{y}_1^T$ is shorthand for all the data from time $t = 1$ to $t = T$. n is the number of data points. The likelihood function comes from the likelihood function for a multivariate normal distribution since $\boldsymbol{X}_t|\boldsymbol{x}_{t-1}$ is multivariate normal and $\boldsymbol{Y}_t|\boldsymbol{x}_t$ is multivariate normal. Here $\boldsymbol{X}_t$ denotes the random variable "hidden states at time t and $\boldsymbol{x}_t$ is a realization from that random variable.

We expand out the terms in the joint log-likelihood to give the rather longer form:

$$\begin{aligned}&\log \mathbf{L}(\boldsymbol{y}_1^T, \boldsymbol{x}_1^T|\Theta) = \\ &- \frac{1}{2}\sum_1^T \Big[(\boldsymbol{y}_t)^\top \mathbf{R}^{-1}\boldsymbol{y}_t - (\boldsymbol{y}_t)^\top \mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t - (\mathbf{Z}\boldsymbol{x}_t)^\top \mathbf{R}^{-1}\boldsymbol{y}_t - \mathbf{a}^\top \mathbf{R}^{-1}\boldsymbol{y}_t \\ &- (\boldsymbol{y}_t)^\top \mathbf{R}^{-1}\mathbf{a} + (\mathbf{Z}\boldsymbol{x}_t)^\top \mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t + \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t + (\mathbf{Z}\boldsymbol{x}_t)^\top \mathbf{R}^{-1}\mathbf{a} \\ &+ \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{a}\Big] - \frac{T}{2}\log|\mathbf{R}| \\ &- \frac{1}{2}\sum_2^T \Big[(\boldsymbol{x}_t)^\top \mathbf{Q}^{-1}\boldsymbol{x}_t - (\boldsymbol{x}_t)^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1} - (\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1}\boldsymbol{x}_t \\ &- \mathbf{u}^\top \mathbf{Q}^{-1}\boldsymbol{x}_t - (\boldsymbol{x}_t)^\top \mathbf{Q}^{-1}\mathbf{u} + (\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1} \\ &+ (\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1}\mathbf{u} + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{u}\Big] - \frac{T-1}{2}\log|\mathbf{Q}| \\ &- \frac{1}{2}\Big[(\boldsymbol{x}_1)^\top (V_1)^{-1}\boldsymbol{x}_1 - \boldsymbol{\xi}^\top (V_1)^{-1}\boldsymbol{x}_1 - (\boldsymbol{x}_1)^\top (V_1)^{-1}\boldsymbol{\xi} + (\boldsymbol{\xi})^\top (V_1)^{-1}\boldsymbol{\xi}\Big] \\ &- \frac{1}{2}\log|V_1| - \frac{n}{2}\log 2\pi\end{aligned} \tag{3}$$

This likelihood looks a little different than that in Shumway and Stoffer (2006) since here, $\boldsymbol{\xi} = \mathrm{E}[\boldsymbol{x}_1]$ not $\mathrm{E}[\boldsymbol{x}_0]$ and thus the second summation is 2 to T rather than 1 to T. Note that all bolded elements are column vectors (lower case) and

[3]This is not the log likelihood output by the Kalman filter and used in model selection. That would be the $\log \mathbf{L}(\boldsymbol{y}_1^T|\Theta)$ equals the marginal or expected log likelihood: $\mathrm{E}_{\mathbf{X}|\mathbf{y}} \log \mathbf{L}(\boldsymbol{y}_1^T, \boldsymbol{x}_1^T|\Theta)$.

matrices (upper case). $\mathbf{A}^\top$ is the transpose of matrix $\mathbf{A}$, $\mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$, and $|\mathbf{A}|$ is the determinant of $\mathbf{A}$. Parameters are non-italic while elements that are slanted are realizations of a random variable ($\boldsymbol{x}$ and $\boldsymbol{y}$ are slated)[4]

2 The EM algorithm

The algorithm cycles iteratively between an expectation step followed by a maximization step.

Expectation step, the expected values of the hidden states conditioned all the data and on a set of parameters at iteration i, $\hat{\Theta}_i$, are computed using the Kalman smoother[5]. The output from the Kalman smoother provides

$$\widetilde{\mathbf{x}}_t = \mathrm{E}_{\mathbf{X}|\mathbf{y}}(\boldsymbol{x}_t|\boldsymbol{y}_1^T, \hat{\Theta}_i) \tag{4a}$$

$$\widetilde{\mathbf{V}}_t = \mathrm{var}(\boldsymbol{X}_t|\boldsymbol{y}_1^T, \hat{\Theta}_i) \tag{4b}$$

$$\widetilde{\mathbf{V}}_{t,t-1} = \mathrm{cov}(\boldsymbol{X}_t, \boldsymbol{X}_{t-1}|\boldsymbol{y}_1^T, \hat{\Theta}_i) \tag{4c}$$

$$\text{From } \widetilde{\mathbf{x}}_t, \widetilde{\mathbf{V}}_t, \text{ and } \widetilde{\mathbf{V}}_{t,t-1}, \text{ we can compute} \tag{4d}$$

$$\widetilde{\mathbf{P}}_t = \mathrm{E}_{\mathbf{X}|\mathbf{y}}(\boldsymbol{x}_t(\boldsymbol{x}_t)^\top|\boldsymbol{y}_1^T, \hat{\Theta}_i) = \widetilde{\mathbf{V}}_t + \widetilde{\mathbf{x}}_t(\widetilde{\mathbf{x}}_t)^\top \tag{4e}$$

$$\widetilde{\mathbf{P}}_{t,t-1} = \mathrm{E}_{\mathbf{X}|\mathbf{y}}(\boldsymbol{x}_t(\boldsymbol{x}_{t-1})^\top|\boldsymbol{y}_1^T, \hat{\Theta}_i) = \widetilde{\mathbf{V}}_{t,t-1} + \widetilde{\mathbf{x}}_t(\widetilde{\mathbf{x}}_{t-1})^\top \tag{4f}$$

The subscript on the expectation, E, denotes that the expectation is taken over the hidden states, $\boldsymbol{X}$, conditioned on the observed data, $\boldsymbol{y}$. The right sides of equations (4e) and (4f) arise from the computational formula for variance and covariance:

$$\begin{aligned}\mathrm{var}(X) &= \mathrm{E}(XX^\top) - \mathrm{E}(X)\,\mathrm{E}(X)\\ \mathrm{cov}(X,Y) &= \mathrm{E}(XY^\top) - \mathrm{E}(X)\,\mathrm{E}(Y)^\top.\end{aligned}$$

Maximization step: a new parameter set $\hat{\Theta}_{i+1}$ is computed by finding the parameters that maximize the *expected* log-likelihood function (see section 2.1) using $\widetilde{\mathbf{x}}_t$, $\widetilde{\mathbf{P}}_t$ and $\widetilde{\mathbf{P}}_{t,t-1}$ from iteration i. The equations that give the parameters for the next iteration $(i+1)$ are called the update equations and most of this appendix is devoted to the derivation of the update equations.

After one iteration of the expectation and maximization steps, the cycle is then repeated. New $\widetilde{\mathbf{x}}_t$, $\widetilde{\mathbf{P}}_t$ and $\widetilde{\mathbf{P}}_{t,t-1}$ are computed using $\hat{\Theta}_{i+1}$, and then a new set of parameters $\hat{\Theta}_{i+2}$ is generated. This cycle is continued until the likelihood no longer increases more than a specified tolerance level. This algorithm is guaranteed to increase in likelihood at each iteration (if it does not, it means

[4] In matrix algebra, a capitol bolded letter indicates a matrix. Unfortunately in statistics, the capitol letter convention is used for random variables. Fortunately, this derivation does not need to reference random variables except indirectly when using expectations. Thus, I use capitols to refer to matrices not random variables. The one exception is the reference to $\boldsymbol{X}$ and in this case a bolded *slanted* capitol is used.

[5] The Kalman smoother gives estimates conditioned on $\boldsymbol{y}_1^T$. It uses the output from the Kalman filter, which gives $\boldsymbol{y}_1^{t-1}$

there is an error in one's update equations). The algorithm must be started from an initial set of parameter values $\hat{\Theta}_1$. The algorithm is not particularly sensitive to the initial conditions but the surface could definitely be multi-modal and have local maxima. See section 5 on using Monte Carlo initialization to ensure that the global maximum is found.

2.1 The expected log-likelihood function

The likelihood function that is maximized in the "M" step is the expected log-likelihood function where the expectation is taken over $(\boldsymbol{X}_1^T|\boldsymbol{y}_1^T)$, meaning the set of all possible hidden states conditioned on all the data. We denote the expected log-likelihood by Ψ. Using the log likelihood equation (3), Ψ is:

$$
\begin{aligned}
&\mathrm{E}_{\mathbf{X}|\mathbf{y}} \log \mathbf{L}(\boldsymbol{y}_1^T, \boldsymbol{x}_1^T|\Theta) = \Psi = \\
&\quad -\frac{1}{2}\sum_1^T \Big((\boldsymbol{y}_t)^\top \mathbf{R}^{-1}\boldsymbol{y}_t - \mathrm{E}_{\boldsymbol{X}|\boldsymbol{y}}[(\boldsymbol{y}_t)^\top \mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t] - \mathrm{E}_{\mathbf{X}|\mathbf{y}}[(\mathbf{Z}\boldsymbol{x}_t)^\top \mathbf{R}^{-1}\boldsymbol{y}_t] \\
&\quad - \mathbf{a}^\top \mathbf{R}^{-1}\boldsymbol{y}_t - (\boldsymbol{y}_t)^\top \mathbf{R}^{-1}\mathbf{a} + \mathrm{E}_{\mathbf{X}|\mathbf{y}}[(\mathbf{Z}\boldsymbol{x}_t)^\top \mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t] \\
&\quad + \mathrm{E}_{\mathbf{X}|\mathbf{y}}[\mathbf{a}^\top \mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t] + \mathrm{E}_{\mathbf{X}|\mathbf{y}}[(\mathbf{Z}\boldsymbol{x}_t)^\top \mathbf{R}^{-1}\mathbf{a}] + \mathbf{a}^\top \mathbf{R}^{-1}\mathbf{a} \Big) - \frac{T}{2}\log|\mathbf{R}| \\
&\quad -\frac{1}{2}\sum_2^T \Big(\mathrm{E}_{\mathbf{X}|\mathbf{y}}[(\boldsymbol{x}_t)^\top \mathbf{Q}^{-1}\boldsymbol{x}_t] - \mathrm{E}_{\mathbf{X}|\mathbf{y}}[(\boldsymbol{x}_t)^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1}] \\
&\quad - \mathrm{E}_{\mathbf{X}|\mathbf{y}}[(\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1}\boldsymbol{x}_t] - \mathrm{E}_{\mathbf{X}|\mathbf{y}}[\mathbf{u}^\top \mathbf{Q}^{-1}\boldsymbol{x}_t] - \mathrm{E}_{\mathbf{X}|\mathbf{y}}[(\boldsymbol{x}_t)^\top \mathbf{Q}^{-1}\mathbf{u}] \\
&\quad + \mathrm{E}_{\mathbf{X}|\mathbf{y}}[(\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1}] + \mathrm{E}_{\mathbf{X}|\mathbf{y}}[\mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1}] \\
&\quad + \mathrm{E}_{\mathbf{X}|\mathbf{y}}[(\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1}\mathbf{u}] + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{u} \Big) - \frac{T-1}{2}\log|\mathbf{Q}| \\
&\quad -\frac{1}{2}\Big(\mathrm{E}_{\mathbf{X}|\mathbf{y}}[(\boldsymbol{x}_1)^\top (V_1)^{-1}\boldsymbol{x}_1] - \mathrm{E}_{\mathbf{X}|\mathbf{y}}[\boldsymbol{\xi}^\top (V_1)^{-1}\boldsymbol{x}_1] \\
&\quad - \mathrm{E}_{\mathbf{X}|\mathbf{y}}[(\boldsymbol{x}_1)^\top (V_1)^{-1}\boldsymbol{\xi}] + (\boldsymbol{\xi})^\top (V_1)^{-1}\boldsymbol{\xi} \Big) - \frac{1}{2}\log|V_1| - \frac{n}{2}\log\pi
\end{aligned}
\tag{5}
$$

We will reference the expected log-likelihood throughout our derivation of the update equations; it could be written more concisely, but for deriving the update equations, we'll keep in this long form. The new parameters for the maximization step are those parameters that maximize the expected log likelihood Ψ. The equations for these new parameters are termed the update equations.

3 The unconstrained update equations

In this section, we show the derivation of the update equations when all elements of a parameter matrix are estimated and are all allowed to be different; these are the update equations one will see in Shumway and Stoffer's text. If some of

the values are fixed or are shared, the derivations are similar but they get more cluttered. Section 3 shows the general update equations when there are fixed or shared values in the parameter matrices. The general update equations are used in the MARSS R package.

To derive the update equations, we will find the parameters values that maximize Ψ (equation 5) by partial differentiation of Ψ with respect to the parameter of interest, and then solve for the parameter value that sets the partial derivative to zero. The partial differentiation is with respect to each individual parameter element, for example each u_j in the vector $\mathbf{u}$. The idea is to single out those terms in equation (5) that involve u_j (say), differentiate by u_j, set this to zero and solve for u_j. This gives the new u_j that maximizes the partial derivative with respect to u_j of the expected log-likelihood. Matrix calculus gives us a way to jointly maximize Ψ with respect to all elements (not just element j) in a parameter vector or matrix.

Deriving the update equations is tedious. However, understanding exactly how to do it is critical if one wants to develop extensions of the linear MARSS model used in our paper. Before commencing, we need some definitions from matrix derivation. The partial derivative of a scalar (Ψ is a scalar) with respect to some column vector $\mathbf{b}$ (which has elements b_1, b_2 . . .) is

$$\frac{\partial \Psi}{\partial \mathbf{b}} = \begin{bmatrix} \frac{\partial \Psi}{\partial b_1} & \frac{\partial \Psi}{\partial b_2} & \cdots & \frac{\partial \Psi}{\partial b_n} \end{bmatrix}$$

Note that the derivative of a column vector $\mathbf{b}$ is a row vector. The partial derivatives of a scalar with respect to some $n \times n$ matrix $\mathbf{B}$ is

$$\frac{\partial \Psi}{\partial \mathbf{b}} = \begin{bmatrix} \frac{\partial \Psi}{\partial b_{1,1}} & \frac{\partial \Psi}{\partial b_{2,1}} & \cdots & \frac{\partial \Psi}{\partial b_{n,1}} \\ \frac{\partial \Psi}{\partial b_{1,2}} & \frac{\partial \Psi}{\partial b_{2,2}} & \cdots & \frac{\partial \Psi}{\partial b_{n,2}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial \Psi}{\partial b_{1,n}} & \frac{\partial \Psi}{\partial b_{2,n}} & \cdots & \frac{\partial \Psi}{\partial b_{n,n}} \end{bmatrix}$$

Note that the indexing is interchanged; $\partial\Psi/\partial b_{i,j} = \left[\partial\Psi/\partial\mathbf{B}\right]_{j,i}$. For $\mathbf{Q}$ and $\mathbf{R}$, this is unimportant because they are variance-covariance matrices and are symmetric. For $\mathbf{B}$ and $\mathbf{Z}$, we must be careful because these may not be symmetric. Table 1 shows matrix differentials that are used in our derivation.

3.1 The update equation for u (unconstrained)

Take the partial derivative of Ψ with respect to $\mathbf{u}$, which is a $m \times 1$ column vector. All parameters other than $\mathbf{u}$ are fixed to constant values (because we are doing partial derivation). Since the derivative of a constant is 0, terms not

Table 1: Derivatives of a scalar with respect to vectors and matrices. In the following $\mathbf{a}$ and $\mathbf{c}$ are $n \times 1$ column vectors, $\mathbf{b}$ and $\mathbf{d}$ are $m \times 1$ column vectors, $\mathbf{D}$ is a $n \times m$ matrix, and $\mathbf{C}$ is a $n \times n$ matrix. Note, all the numerators in the differentials reduce to scalars. Both the vectorized and non-vectorized versions are shown; vec is defined at the bottom of the table.

$$\partial(\mathbf{a}^\top \mathbf{c})/\partial \mathbf{a} = \partial(\mathbf{c}^\top \mathbf{a})/\partial \mathbf{a} = \mathbf{c}^\top \tag{6}$$

$$\begin{aligned} &\partial(\mathbf{a}^\top \mathbf{D}\mathbf{b})/\partial \mathbf{D} = \partial(\mathbf{b}^\top \mathbf{D}^\top \mathbf{a})/\partial \mathbf{D} = \mathbf{b}\mathbf{a}^\top \\ &\partial(\mathbf{a}^\top \mathbf{D}\mathbf{b})/\partial \operatorname{vec}(\mathbf{D}) = \partial(\mathbf{b}^\top \mathbf{D}^\top \mathbf{a})/\partial \operatorname{vec}(\mathbf{D}) = \big(\operatorname{vec}(\mathbf{b}\mathbf{a}^\top)\big)^\top \end{aligned} \tag{7}$$

$$\begin{aligned} &\partial(\log|\mathbf{C}|)/\partial \mathbf{C} = -\partial(\log|\mathbf{C}^{-1}|)/\partial \mathbf{C} = (\mathbf{C}^\top)^{-1} = \mathbf{C}^{-\top} \\ &\quad = \mathbf{C}^{-1} \text{ if } \mathbf{C} \text{ is symmetric} \\ &\partial(\log|\mathbf{C}|)/\partial \operatorname{vec}(\mathbf{C}) = \big(\operatorname{vec}(\mathbf{C}^{-\top})\big)^\top \end{aligned} \tag{8}$$

$$\begin{aligned} &\partial(\mathbf{b}^\top \mathbf{D}^\top \mathbf{C}\mathbf{D}\mathbf{d})/\partial \mathbf{D} = \mathbf{d}\mathbf{b}^\top \mathbf{D}^\top \mathbf{C} + \mathbf{b}\mathbf{d}^\top \mathbf{D}^\top \mathbf{C}^\top \\ &\partial(\mathbf{b}^\top \mathbf{D}^\top \mathbf{C}\mathbf{D}\mathbf{d})/\partial \operatorname{vec}(\mathbf{D}) = \big(\operatorname{vec}(\mathbf{d}\mathbf{b}^\top \mathbf{D}^\top \mathbf{C} + \mathbf{b}\mathbf{d}^\top \mathbf{D}^\top \mathbf{C}^\top)\big)^\top \\ &\text{If } \mathbf{b} = \mathbf{d} \text{ and } \mathbf{C} \text{ is symmetric then the sum reduces to } 2\mathbf{b}\mathbf{b}^\top \mathbf{D}^\top \mathbf{C} \end{aligned} \tag{9}$$

$$\partial(\mathbf{a}^\top \mathbf{C}\mathbf{a})/\partial \mathbf{a} = \partial(\mathbf{a}\mathbf{C}^\top \mathbf{a}^\top)/\partial \mathbf{a} = 2\mathbf{a}^\top \mathbf{C} \tag{10}$$

$$\begin{aligned} &\partial(\mathbf{a}^\top \mathbf{C}^{-1}\mathbf{c})/\partial \mathbf{C} = -\mathbf{C}^{-1}\mathbf{a}\mathbf{c}^\top \mathbf{C}^{-1} \\ &\partial(\mathbf{a}^\top \mathbf{C}^{-1}\mathbf{c})/\partial \operatorname{vec}(\mathbf{C}) = -\big(\operatorname{vec}(\mathbf{C}^{-1}\mathbf{a}\mathbf{c}^\top \mathbf{C}^{-1})\big)^\top \end{aligned} \tag{11}$$

$$\partial f/\partial \mathbf{Z} = \frac{\partial f}{\partial \boldsymbol{Y}} \frac{\partial \boldsymbol{Y}}{\partial \mathbf{Z}} \text{ the chain rule} \tag{12}$$

$$\operatorname{vec}(\mathbf{D}_{n,m}) \equiv \begin{bmatrix} d_{1,1} \\ \cdots \\ d_{n,1} \\ d_{1,2} \\ \cdots \\ d_{n,2} \\ \cdots \\ d_{1,m} \\ \cdots \\ d_{n,m} \end{bmatrix}$$

$\mathbf{C}^{-1} \equiv$ inverse of $\mathbf{C}$ $\quad \mathbf{C}^{-\top} = \big(\mathbf{C}^{-1}\big)^\top = \big(\mathbf{C}^\top\big)^{-1}$
$\mathbf{D}^\top \equiv$ transpose of $\mathbf{D}$ $\quad |\mathbf{C}| \equiv$ determinant of $\mathbf{C}$

involving $\mathbf{u}$ will equal 0 and drop out. The subscript, $\boldsymbol{X}|\boldsymbol{y}$, on the expectation, E, has been dropped to remove clutter. Taking the derivative to equation (5) with respect to $\mathbf{u}$:

$$\begin{aligned}\partial\Psi/\partial\mathbf{u} = -\frac{1}{2}\sum_{t=2}^{T}\Big(&- \mathrm{E}[\partial((\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{u}] - \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\boldsymbol{x}_t)/\partial\mathbf{u}] \\ &+ \mathrm{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{u}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1})/\partial\mathbf{u}] \\ &+ \partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{u}\Big)\end{aligned} \tag{13}$$

Using relations (6) and (10) and using $\mathbf{Q}^{-1} = (\mathbf{Q}^{-1})^\top$, we have

$$\begin{aligned}\partial\Psi/\partial\mathbf{u} = -\frac{1}{2}\sum_{t=2}^{T}\Big(&- \mathrm{E}[(\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}] - \mathrm{E}[(\mathbf{Q}^{-1}\boldsymbol{x}_t)^\top] \\ &+ \mathrm{E}[(\mathbf{B}\boldsymbol{x}_{t-1})^\top\mathbf{Q}^{-1}] + \mathrm{E}[(\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1})^\top] + 2\mathbf{u}^\top\mathbf{Q}^{-1}\Big)\end{aligned} \tag{14}$$

The parameters can be pulled out of the expectations[6] and the $-1/2$ removed, giving

$$\partial\Psi/\partial\mathbf{u} = \sum_{t=2}^{T}\big(\mathrm{E}[(\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}] - \mathrm{E}[(\boldsymbol{x}_{t-1})^\top]\mathbf{B}^\top\mathbf{Q}^{-1} - \mathbf{u}^\top\mathbf{Q}^{-1}\big) \tag{15}$$

Set the left side to zero (a $1 \times m$ matrix of zeros) and transpose the whole equation. $\mathbf{Q}^{-1}$ cancels out[7] by multiplying on the left by $\mathbf{Q}$ (left since we just transposed the whole equation), giving

$$\mathbf{0} = \sum_{t=2}^{T}\big(\mathrm{E}[\boldsymbol{x}_t] - \mathbf{B}\,\mathrm{E}[\boldsymbol{x}_{t-1}] - \mathbf{u}\big) = \sum_{t=2}^{T}\big(\mathrm{E}[\boldsymbol{x}_t] - \mathbf{B}\,\mathrm{E}[\boldsymbol{x}_{t-1}]\big) - (T-1)\mathbf{u} \tag{16}$$

Solving for $\mathbf{u}$ and replacing the expectations with the Kalman smoother output, gives us the new $\mathbf{u}$ that maximizes Ψ,

$$\mathbf{u}_{\text{new}} = \frac{1}{T-1}\sum_{t=2}^{T}\big(\widetilde{\mathbf{x}}_t - \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\big) \tag{17}$$

[6]The expectation is an integral over $\boldsymbol{x}$ and the parameters are not functions of $\boldsymbol{x}$ so they can be pulled out of the expectations.

[7]$\mathbf{Q}$ is a variance-covariance matrix and is invertable. $\mathbf{Q}^{-1}\mathbf{Q} = \mathbf{I}$, the identity matrix.

3.2 The update equation for B (unconstrained)

Take the derivative of Ψ with respect to $\mathbf{B}$. Terms not involving $\mathbf{B}$, equal 0 and drop out.

$$\begin{aligned}
\partial\Psi/\partial\mathbf{B} &= -\frac{1}{2}\sum_{t=2}^{T}\Big(- \mathrm{E}[\partial((\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1})/\partial\mathbf{B}] \\
&\quad - \mathrm{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top\mathbf{Q}^{-1}\boldsymbol{x}_t)/\partial\mathbf{B}] + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top\mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{x}_{t-1}))/\partial\mathbf{B}] \\
&\quad + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{B}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1})/\partial\mathbf{B}]\Big) \\
&= -\frac{1}{2}\sum_{t=2}^{T}\Big(- \mathrm{E}[\partial((\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1})/\partial\mathbf{B}] \\
&\quad - \mathrm{E}[\partial((\boldsymbol{x}_{t-1})^\top\mathbf{B}^\top\mathbf{Q}^{-1}\boldsymbol{x}_t)/\partial\mathbf{B}] + \mathrm{E}[\partial((\boldsymbol{x}_{t-1})^\top\mathbf{B}^\top\mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{x}_{t-1}))/\partial\mathbf{B}] \\
&\quad + \mathrm{E}[\partial((\boldsymbol{x}_{t-1})^\top\mathbf{B}^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\mathbf{B}] + \mathrm{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1})/\partial\mathbf{B}]\Big)
\end{aligned} \tag{18}$$

Using relations (7) and (9), we have

$$\begin{aligned}
\partial\Psi/\partial\mathbf{B} = -\frac{1}{2}\sum_{t=2}^{T}\Big(&- \mathrm{E}[\boldsymbol{x}_{t-1}(\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}] - \mathrm{E}[\boldsymbol{x}_{t-1}(\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}] \\
&+ 2\,\mathrm{E}[\boldsymbol{x}_{t-1}(\boldsymbol{x}_{t-1})^\top\mathbf{B}^\top\mathbf{Q}^{-1}] + \mathrm{E}[\boldsymbol{x}_{t-1}\mathbf{u}^\top\mathbf{Q}^{-1}] + \mathrm{E}[\boldsymbol{x}_{t-1}\mathbf{u}^\top\mathbf{Q}^{-1}]\Big)
\end{aligned} \tag{19}$$

Pulling the parameters out of the expectations and using $\mathbf{Q}^{-1} = (\mathbf{Q}^{-1})^\top$, we have

$$\begin{aligned}
\partial\Psi/\partial\mathbf{B} = -\frac{1}{2}\sum_{t=2}^{T}\Big(&- 2\,\mathrm{E}[\boldsymbol{x}_{t-1}(\boldsymbol{x}_t)^\top]\mathbf{Q}^{-1} \\
&+ 2\,\mathrm{E}[\boldsymbol{x}_{t-1}(\boldsymbol{x}_{t-1})^\top]\mathbf{B}^\top\mathbf{Q}^{-1} + 2\,\mathrm{E}[\boldsymbol{x}_{t-1}\mathbf{u}^\top\mathbf{Q}^{-1}]\Big)
\end{aligned} \tag{20}$$

Set the left side to zero (an $m \times m$ matrix of zeros), cancel out $\mathbf{Q}^{-1}$ by multiplying by $\mathbf{Q}$ on the right, get rid of the -1/2, and transpose the whole equation to give

$$\begin{aligned}
\mathbf{0} &= \sum_{t=2}^{T}\big(\mathrm{E}[\boldsymbol{x}_t(\boldsymbol{x}_{t-1})^\top] - \mathbf{B}\,\mathrm{E}[\boldsymbol{x}_{t-1}(\boldsymbol{x}_{t-1})^\top] - \mathbf{u}\,\mathrm{E}[(\boldsymbol{x}_{t-1})^\top]\big) \\
&= \sum_{t=2}^{T}\big(\widetilde{\mathbf{P}}_{t,t-1} - \mathbf{B}\widetilde{\mathbf{P}}_{t-1} - \mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top\big)
\end{aligned} \tag{21}$$

The last line replaced the expectations with the Kalman smoother output from equation (4). Solving for $\mathbf{B}$ and noting that $\widetilde{\mathbf{P}}_{t-1}$ is like a variance-covariance

matrix and is invertable, gives us the new $\mathbf{B}$ that maximizes Ψ,

$$\mathbf{B}_{\text{new}} = \Big(\sum_{t=2}^{T} \big(\widetilde{\mathbf{P}}_{t,t-1} - \mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top \big) \Big) \Big(\sum_{t=2}^{T} \widetilde{\mathbf{P}}_{t-1} \Big)^{-1} \tag{22}$$

Because all the equations above also apply to block-diagonal matrices, the derivation immediately generalizes to the case where $\mathbf{B}$ is an unconstrained block diagonal matrix:

$$\mathbf{B} = \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} & 0 & 0 & 0 & 0 & 0 \\ b_{2,1} & b_{2,2} & b_{2,3} & 0 & 0 & 0 & 0 & 0 \\ b_{3,1} & b_{3,2} & b_{3,3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_{4,4} & b_{4,5} & 0 & 0 & 0 \\ 0 & 0 & 0 & b_{5,4} & b_{5,5} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & b_{6,6} & b_{6,7} & b_{6,8} \\ 0 & 0 & 0 & 0 & 0 & b_{7,6} & b_{7,7} & b_{7,8} \\ 0 & 0 & 0 & 0 & 0 & b_{8,6} & b_{8,7} & b_{8,8} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 & 0 & 0 \\ 0 & \mathbf{B}_2 & 0 \\ 0 & 0 & \mathbf{B}_3 \end{bmatrix}$$

For the block diagonal $\mathbf{B}$,

$$\mathbf{B}_{i,\text{new}} = \Big(\sum_{t=2}^{T} \big(\widetilde{\mathbf{P}}_{t,t-1} - \mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top \big) \Big)_i \Big(\sum_{t=2}^{T} \widetilde{\mathbf{P}}_{t-1} \Big)_i^{-1} \tag{23}$$

where the subscript i means we take the parts of the matrices that are analogous to $\mathbf{B}_i$; take the whole part within the parentheses not the individual matrices inside the parentheses) . If $\mathbf{B}_i$ is comprised of rows a to b and columns c to d of matrix $\mathbf{B}$, then we take rows a to b and columns c to d of matrices subscripted by i in equation (23).

3.3 The update equation for Q (unconstrained)

The usual way to do this derivation is to use what is known as the "trace trick" which will pull the $\mathbf{Q}^{-1}$ out to the left of the $\mathbf{c}^\top \mathbf{Q}^{-1} \mathbf{b}$ terms which appear in the likelihood (5). Here I'm showing a less elegant derivation that plods step by step through each of the likelihood terms. Take the derivative of Ψ with respect to $\mathbf{Q}$. Terms not involving $\mathbf{Q}$ equal 0 and drop out.

$$\begin{aligned} \partial\Psi/\partial\mathbf{Q} = -\frac{1}{2} \sum_{t=2}^{T} \Big(& \mathrm{E}[\partial((\boldsymbol{x}_t)^\top \mathbf{Q}^{-1} \boldsymbol{x}_t)/\partial\mathbf{Q}] - \mathrm{E}[\partial((\boldsymbol{x}_t)^\top \mathbf{Q}^{-1} \mathbf{B} \boldsymbol{x}_{t-1})/\partial\mathbf{Q}] \\ & - \mathrm{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1} \boldsymbol{x}_t)/\partial\mathbf{Q}] - \mathrm{E}[\partial((\boldsymbol{x}_t)^\top \mathbf{Q}^{-1} \mathbf{u})/\partial\mathbf{Q}] \\ & - \mathrm{E}[\partial(\mathbf{u}^\top \mathbf{Q}^{-1} \boldsymbol{x}_t)/\partial\mathbf{Q}] + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1} \mathbf{B} \boldsymbol{x}_{t-1})/\partial\mathbf{Q}] \\ & + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1} \mathbf{u})/\partial\mathbf{Q}] + \mathrm{E}[\partial(\mathbf{u}^\top \mathbf{Q}^{-1} \mathbf{B} \boldsymbol{x}_{t-1})/\partial\mathbf{Q}] \\ & + \partial(\mathbf{u}^\top \mathbf{Q}^{-1} \mathbf{u})/\partial\mathbf{Q} \Big) - \partial\Big(\frac{T-1}{2} \log|\mathbf{Q}| \Big)/\partial\mathbf{Q} \end{aligned} \tag{24}$$

We use relations (11) and (8) to do the differentiation. Notice that all the terms in the summation are of the form $\mathbf{c}^\top \mathbf{Q}^{-1}\mathbf{b}$, and thus we group all the $\mathbf{c}^\top\mathbf{b}$ inside one set of parentheses. Also there is a minus that comes from equation (11) and it cancels out the minus in front of the initial $-1/2$.

$$\begin{aligned}\partial\Psi/\partial\mathbf{Q} = \frac{1}{2}\sum_{t=2}^{T}\mathbf{Q}^{-1}\Big(&\operatorname{E}[\boldsymbol{x}_t(\boldsymbol{x}_t)^\top] - \operatorname{E}[\boldsymbol{x}_t(\mathbf{B}\boldsymbol{x}_{t-1})^\top] - \operatorname{E}[\mathbf{B}\boldsymbol{x}_{t-1}(\boldsymbol{x}_t)^\top] \\ &- \operatorname{E}[\boldsymbol{x}_t\mathbf{u}^\top] - \operatorname{E}[\mathbf{u}(\boldsymbol{x}_t)^\top] + \operatorname{E}[\mathbf{B}\boldsymbol{x}_{t-1}(\mathbf{B}\boldsymbol{x}_{t-1})^\top] + \operatorname{E}[\mathbf{B}\boldsymbol{x}_{t-1}\mathbf{u}^\top] \\ &+ \operatorname{E}[\mathbf{u}(\mathbf{B}\boldsymbol{x}_{t-1})^\top] + \mathbf{u}\mathbf{u}^\top\Big)\mathbf{Q}^{-1} - \frac{T-1}{2}\mathbf{Q}^{-1}\end{aligned} \tag{25}$$

Pulling the parameters out of the expectations and using $(\mathbf{B}\boldsymbol{x}_t)^\top = (\boldsymbol{x}_t)^\top\mathbf{B}^\top$, we have

$$\begin{aligned}\partial\Psi/\partial\mathbf{Q} = \frac{1}{2}\sum_{t=2}^{T}\mathbf{Q}^{-1}\Big(&\operatorname{E}[\boldsymbol{x}_t(\boldsymbol{x}_t)^\top] - \operatorname{E}[\boldsymbol{x}_t(\boldsymbol{x}_{t-1})^\top]\mathbf{B}^\top - \mathbf{B}\operatorname{E}[\boldsymbol{x}_{t-1}(\boldsymbol{x}_t)^\top] \\ &- \operatorname{E}[\boldsymbol{x}_t]\mathbf{u}^\top - \mathbf{u}\operatorname{E}[(\boldsymbol{x}_t)^\top] + \mathbf{B}\operatorname{E}[\boldsymbol{x}_{t-1}(\boldsymbol{x}_{t-1})^\top]\mathbf{B}^\top + \mathbf{B}\operatorname{E}[\boldsymbol{x}_{t-1}]\mathbf{u}^\top \\ &+ \mathbf{u}\operatorname{E}[(\boldsymbol{x}_{t-1})^\top]\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\Big)\mathbf{Q}^{-1} - \frac{T-1}{2}\mathbf{Q}^{-1}\end{aligned} \tag{26}$$

We rewrite the partial derivative in terms of the Kalman smoother output:

$$\begin{aligned}\partial\Psi/\partial\mathbf{Q} = \frac{1}{2}\sum_{t=2}^{T}\mathbf{Q}^{-1}\Big(&\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}(\widetilde{\mathbf{x}}_t)^\top \\ &+ \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top\mathbf{B}^\top \\ &+ \mathbf{u}\mathbf{u}^\top\Big)\mathbf{Q}^{-1} - \frac{T-1}{2}\mathbf{Q}^{-1}\end{aligned} \tag{27}$$

Setting this to zero (a $m \times m$ matrix of zeros), we cancel out $\mathbf{Q}^{-1}$ by multiplying by $\mathbf{Q}$ twice, once on the left and once on the right and get rid of the $1/2$.

$$\begin{aligned}\mathbf{0} = \sum_{t=2}^{T}\Big(&\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}(\widetilde{\mathbf{x}}_t)^\top \\ &+ \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\Big) - \mathbf{Q}(T-1)\end{aligned} \tag{28}$$

We can then solve for $\mathbf{Q}$, giving us the new $\mathbf{Q}$ that maximizes Ψ,

$$\begin{aligned}\mathbf{Q}_{\text{new}} = \frac{1}{T-1}\sum_{t=2}^{T}\Big(&\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}(\widetilde{\mathbf{x}}_t)^\top \\ &+ \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\Big)\end{aligned} \tag{29}$$

This derivation immediately generalizes to the case where $\mathbf{Q}$ is a block diagonal matrix:

$$\mathbf{Q} = \begin{bmatrix} q_{1,1} & q_{1,2} & q_{1,3} & 0 & 0 & 0 & 0 & 0 \\ q_{1,2} & q_{2,2} & q_{2,3} & 0 & 0 & 0 & 0 & 0 \\ q_{1,3} & q_{2,3} & q_{3,3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & q_{4,4} & q_{4,5} & 0 & 0 & 0 \\ 0 & 0 & 0 & q_{4,5} & q_{5,5} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & q_{6,6} & q_{6,7} & q_{6,8} \\ 0 & 0 & 0 & 0 & 0 & q_{6,7} & q_{7,7} & q_{7,8} \\ 0 & 0 & 0 & 0 & 0 & q_{6,8} & q_{7,8} & q_{8,8} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1 & 0 & 0 \\ 0 & \mathbf{Q}_2 & 0 \\ 0 & 0 & \mathbf{Q}_3 \end{bmatrix}$$

In this case,

$$\begin{aligned} \mathbf{Q}_{i,\text{new}} = \frac{1}{T-1} \sum_{t=2}^{T} \Big(& \widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}(\widetilde{\mathbf{x}}_t)^\top \\ & + \mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top \Big)_i \end{aligned} \tag{30}$$

where the subscript i means we take the matrix (in the big parentheses) that are analogous to $\mathbf{Q}_i$; take the whole part within the parentheses not the individual matrices inside the parentheses). If $\mathbf{Q}_i$ is comprised of rows a to b and columns c to d of matrix $\mathbf{Q}$, then we take rows a to b and columns c to d of matrices subscripted by i in equation (30).

By the way, $\mathbf{Q}$ is never really unconstrained since it is a variance-covariance matrix and the upper and lower triangles are shared. However, because the shared values are only the symmetric values in the matrix, the derivation still works even though it's technically incorrect (Henderson and Searle, 1979). The constrained update equation for $\mathbf{Q}$ explicitly deals with the shared lower and upper triangles.

3.4 Update equation for a (unconstrained)

Take the derivative of Ψ with respect to $\mathbf{a}$, where $\mathbf{a}$ is a $n \times 1$ column vector. Terms not involving $\mathbf{a}$, equal 0 and drop out.

$$\begin{aligned} \partial\Psi/\partial\mathbf{a} = -\frac{1}{2} \sum_{t=1}^{T} \Big(& - \partial((\boldsymbol{y}_t)^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{a} - \partial(\mathbf{a}^\top\mathbf{R}^{-1}\boldsymbol{y}_t)/\partial\mathbf{a} \\ & + \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{x}_t)^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{a}] + \mathrm{E}[\partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t)/\partial\mathbf{a}] + \partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{a} \Big) \end{aligned} \tag{31}$$

Using relations (6) and (10) and using $\mathbf{R}^{-1} = (\mathbf{R}^{-1})^\top$, we have

$$\begin{aligned} \partial\Psi/\partial\mathbf{a} = -\frac{1}{2} \sum_{t=1}^{T} \Big(& - (\boldsymbol{y}_t)^\top\mathbf{R}^{-1} - (\mathbf{R}^{-1}\boldsymbol{y}_t)^\top + \mathrm{E}[(\mathbf{Z}\boldsymbol{x}_t)^\top\mathbf{R}^{-1}] \\ & + \mathrm{E}[(\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t)^\top] + 2\mathbf{a}^\top\mathbf{R}^{-1} \Big) \end{aligned} \tag{32}$$

Pull the parameters out of the expectations, use $(\mathbf{ab})^\top = \mathbf{b}^\top\mathbf{a}^\top$ and $\mathbf{R}^{-1} = (\mathbf{R}^{-1})^\top$ where needed, and remove the $-1/2$ to get

$$\partial\Psi/\partial\mathbf{a} = \sum_{t=1}^{T}\Big((\boldsymbol{y}_t)^\top\mathbf{R}^{-1} - \mathrm{E}[(\boldsymbol{x}_t)^\top]\mathbf{Z}^\top\mathbf{R}^{-1} - \mathbf{a}^\top\mathbf{R}^{-1}\Big) \tag{33}$$

Set the left side to zero (a $1\times n$ matrix of zeros), take the transpose, and cancel out $\mathbf{R}^{-1}$ by multiplying by $\mathbf{R}$, giving

$$\mathbf{0} = \sum_{t=1}^{T-1}\big(\boldsymbol{y}_t - \mathbf{Z}\,\mathrm{E}[\boldsymbol{x}_t] - \mathbf{a}\big) = \sum_{t=1}^{T-1}\big(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a}\big) \tag{34}$$

Solving for $\mathbf{a}$ gives us the update equation for $\mathbf{a}$:

$$\mathbf{a}_{\mathrm{new}} = \frac{1}{T}\sum_{t=1}^{T-1}\big(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t\big) \tag{35}$$

If the i-th value of $\boldsymbol{y}$ is missing at time t, that would be $y_{i,t}$, then the i-th value of $\mathbf{a}$ from the previous iteration of the EM algorithm, $\mathbf{a}_{i,\mathrm{old}}$, is used in place of the i-th value of $(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t)$ in the summation at time t.

3.5 The update equation for Z (unconstrained)

Take the derivative of Ψ with respect to $\mathbf{Z}$. Terms not involving $\mathbf{Z}$, equal 0 and drop out.

$$\begin{aligned}
&\partial\Psi/\partial\mathbf{Z} = (\text{note } \partial\mathbf{Z} \text{ is } m\times n \text{ while } \mathbf{Z} \text{ is } n\times m)\\
&\quad -\frac{1}{2}\sum_{t=1}^{T}\Big(-\mathrm{E}[\partial((\boldsymbol{y}_t)^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t)/\partial\mathbf{Z}]\\
&\quad -\mathrm{E}[\partial((\mathbf{Z}\boldsymbol{x}_t)^\top\mathbf{R}^{-1}\boldsymbol{y}_t)/\partial\mathbf{Z}] + \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{x}_t)^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t)/\partial\mathbf{Z}]\\
&\quad +\mathrm{E}[\partial((\mathbf{Z}\boldsymbol{x}_t)^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{Z}] + \mathrm{E}[\partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t)/\partial\mathbf{B}]\Big)\\
&= -\frac{1}{2}\sum_{t=1}^{T}\Big(-\mathrm{E}[\partial((\boldsymbol{y}_t)^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t)/\partial\mathbf{Z}]\\
&\quad -\mathrm{E}[\partial((\boldsymbol{x}_t)^\top\mathbf{Z}^\top\mathbf{R}^{-1}\boldsymbol{y}_t)/\partial\mathbf{Z}] + \mathrm{E}[\partial((\boldsymbol{x}_t)^\top\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t)/\partial\mathbf{Z}]\\
&\quad +\mathrm{E}[\partial((\boldsymbol{x}_t)^\top\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{Z}] + \mathrm{E}[\partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t)/\partial\mathbf{Z}]\Big)
\end{aligned} \tag{36}$$

Using relations (7) and (9) and using $\mathbf{R}^{-1} = (\mathbf{R}^{-1})^\top$, we get

$$\begin{aligned}
\partial\Psi/\partial\mathbf{Z} = -\frac{1}{2}\sum_{t=1}^{T}\Big(&-\mathrm{E}[\boldsymbol{x}_t(\boldsymbol{y}_t)^\top\mathbf{R}^{-1}] - \mathrm{E}[\boldsymbol{x}_t(\boldsymbol{y}_t)^\top\mathbf{R}^{-1}]\\
&+2\,\mathrm{E}[\boldsymbol{x}_t(\boldsymbol{x}_t)^\top\mathbf{Z}^\top\mathbf{R}^{-1}] + \mathrm{E}[\boldsymbol{x}_{t-1}\mathbf{a}^\top\mathbf{R}^{-1}] + \mathrm{E}[\boldsymbol{x}_t\mathbf{a}^\top\mathbf{R}^{-1}]\Big)
\end{aligned} \tag{37}$$

Pulling the parameters and $\boldsymbol{y}$ out of the expectations, we have

$$\begin{aligned}\partial\Psi/\partial\mathbf{Z} = -\frac{1}{2}\sum_{t=1}^{T}\Big(&- 2\,\mathrm{E}[\boldsymbol{x}_t](\boldsymbol{y}_t)^\top\mathbf{R}^{-1} + 2\,\mathrm{E}[\boldsymbol{x}_t(\boldsymbol{x}_t)^\top]\mathbf{Z}^\top\mathbf{R}^{-1} \\ &+ 2\,\mathrm{E}[\boldsymbol{x}_t]\mathbf{a}^\top\mathbf{R}^{-1}\Big)\end{aligned} \tag{38}$$

Set the left side to zero (a $m \times n$ matrix of zeros), transpose it all, get rid of the $-1/2$, and cancel out $\mathbf{R}^{-1}$ by multiplying by $\mathbf{R}$ on the left, to give

$$\begin{aligned}\mathbf{0} &= \sum_{t=1}^{T}\left(\boldsymbol{y}_t\,\mathrm{E}[(\boldsymbol{x}_t)^\top] - \mathbf{Z}\,\mathrm{E}[\boldsymbol{x}_t(\boldsymbol{x}_t)^\top] - \mathbf{a}\,\mathrm{E}[(\boldsymbol{x}_t)^\top]\right) \\ &= \sum_{t=1}^{T}\left(\boldsymbol{y}_t(\widetilde{\mathbf{x}}_t)^\top - \mathbf{Z}\widetilde{\mathbf{P}}_t - \mathbf{a}(\widetilde{\mathbf{x}}_t)^\top\right)\end{aligned} \tag{39}$$

Solving for $\mathbf{Z}$ and noting that $\widetilde{\mathbf{P}}_t$ is invertable, gives us the new $\mathbf{Z}$ that maximizes Ψ,

$$\mathbf{Z}_{\text{new}} = \left(\sum_{t=1}^{T}\left((\boldsymbol{y}_t - \mathbf{a})(\widetilde{\mathbf{x}}_t)^\top\right)\right)\left(\sum_{t=1}^{T}\widetilde{\mathbf{P}}_t\right)^{-1} \tag{40}$$

If there are missing values in the data, then rewrite equation (40) as follows:

$$\mathbf{Z}_{\text{new}} = \sum_{t=1}^{T}\left(\mathbf{P}_{\text{inv}}(\boldsymbol{y}_t - \mathbf{a})(\widetilde{\mathbf{x}}_t)^\top\right) = \sum_{t=1}^{T}\mathbf{H}_t \tag{41}$$

where $\mathbf{P}_{\text{inv}} = \left(\sum_{t=1}^{T}\widetilde{\mathbf{P}}_t\right)^{-1}$ and $\mathbf{H}_t$ denotes $\mathbf{P}_{\text{inv}}(\boldsymbol{y}_t - \mathbf{a})(\widetilde{\mathbf{x}}_t)^\top$. If the i-th value of $\boldsymbol{y}$ is missing at time t, $y_{i,t}$, the i-th row of $\mathbf{Z}$ from the previous iteration of the EM algorithm, $\mathbf{Z}_{\text{old}}$, is used in place of the i-th row of $\mathbf{H}_t$ in the summation at time t.

3.6 The update equation for R (unconstrained)

Take the derivative of Ψ with respect to $\mathbf{R}$. Terms not involving $\mathbf{R}$, equal 0 and drop out.

$$\begin{aligned}\partial\Psi/\partial\mathbf{R} = -\frac{1}{2}\sum_{t=1}^{T}\Big(&\mathrm{E}[\partial((\boldsymbol{y}_t)^\top\mathbf{R}^{-1}\boldsymbol{y}_t)/\partial\mathbf{R}] - \mathrm{E}[\partial((\boldsymbol{y}_t)^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t)/\partial\mathbf{R}] \\ &- \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{x}_t)^\top\mathbf{R}^{-1}\boldsymbol{y}_t)/\partial\mathbf{R}] - \mathrm{E}[\partial((\boldsymbol{y}_t)^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{R}] \\ &- \mathrm{E}[\partial(\mathbf{a}^\top\mathbf{R}^{-1}\boldsymbol{y}_t)/\partial\mathbf{R}] + \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{x}_t)^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t)/\partial\mathbf{R}] \\ &+ \mathrm{E}[\partial((\mathbf{Z}\boldsymbol{x}_t)^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{R}] + \mathrm{E}[\partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{x}_t)/\partial\mathbf{R}] \\ &+ \partial(\mathbf{a}^\top\mathbf{R}^{-1}\mathbf{a})/\partial\mathbf{R}\Big) - \partial\big(\frac{T}{2}\log|\mathbf{R}|\big)/\partial\mathbf{R}\end{aligned} \tag{42}$$

We use relations (11) and (8) to do the differentiation. Notice that all the terms in the summation are of the form $\mathbf{c}^\top \mathbf{R}^{-1}\mathbf{b}$, and thus we group all the $\mathbf{c}^\top\mathbf{b}$ inside one set of parentheses. Also there is a minus that comes from equation (11) and cancels out the minus in front of $-1/2$.

$$\begin{aligned}\partial\Psi/\partial\mathbf{R} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{R}^{-1}\Big(&\operatorname{E}[\boldsymbol{y}_t(\boldsymbol{y}_t)^\top] - \operatorname{E}[\boldsymbol{y}_t(\mathbf{Z}\boldsymbol{x}_t)^\top] - \operatorname{E}[\mathbf{Z}\boldsymbol{x}_t(\boldsymbol{y}_t)^\top] \\ &- \operatorname{E}[\boldsymbol{y}_t\mathbf{a}^\top] - \operatorname{E}[\mathbf{a}(\boldsymbol{y}_t)^\top] + \operatorname{E}[\mathbf{Z}\boldsymbol{x}_t(\mathbf{Z}\boldsymbol{x}_t)^\top] + \operatorname{E}[\mathbf{Z}\boldsymbol{x}_t\mathbf{a}^\top] + \operatorname{E}[\mathbf{a}(\mathbf{Z}\boldsymbol{x}_t)^\top] \\ &+ \mathbf{a}\mathbf{a}^\top\Big)\mathbf{R}^{-1} - \frac{T}{2}\mathbf{R}^{-1}\end{aligned} \tag{43}$$

Pulling the parameters and $\boldsymbol{y}$ out of the expectations and using $(\mathbf{Z}\boldsymbol{y}_t)^\top = (\boldsymbol{y}_t)^\top\mathbf{Z}^\top$, we have

$$\begin{aligned}\partial\Psi/\partial\mathbf{R} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{R}^{-1}\Big(&\boldsymbol{y}_t(\boldsymbol{y}_t)^\top - \boldsymbol{y}_t\operatorname{E}[(\boldsymbol{x}_t)^\top]\mathbf{Z}^\top - \mathbf{Z}\operatorname{E}[\boldsymbol{x}_t](\boldsymbol{y}_t)^\top - \boldsymbol{y}_t\mathbf{a}^\top \\ &- \mathbf{a}(\boldsymbol{y}_t)^\top + \mathbf{Z}\operatorname{E}[\boldsymbol{x}_t(\boldsymbol{x}_t)^\top]\mathbf{Z}^\top + \mathbf{Z}\operatorname{E}[\boldsymbol{x}_t]\mathbf{a}^\top + \mathbf{a}\operatorname{E}[(\boldsymbol{x}_t)^\top]\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big)\mathbf{R}^{-1} \\ &- \frac{T}{2}\mathbf{R}^{-1}\end{aligned} \tag{44}$$

We rewrite the partial derivative in terms of the Kalman smoother output:

$$\begin{aligned}\partial\Psi/\partial\mathbf{R} = \frac{1}{2}\sum_{t=1}^{T}\mathbf{R}^{-1}\Big(&\boldsymbol{y}_t(\boldsymbol{y}_t)^\top - \boldsymbol{y}_t(\widetilde{\mathbf{x}}_t)^\top\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{x}}_t(\boldsymbol{y}_t)^\top - \boldsymbol{y}_t\mathbf{a}^\top - \mathbf{a}(\boldsymbol{y}_t)^\top \\ &+ \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}(\widetilde{\mathbf{x}}_t)^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big)\mathbf{R}^{-1} - \frac{T}{2}\mathbf{R}^{-1}\end{aligned} \tag{45}$$

Setting this to zero (a $n \times n$ matrix of zeros), we cancel out $\mathbf{R}^{-1}$ by multiplying by $\mathbf{R}$ twice, once on the left and once on the right, and get rid of the $1/2$.

$$\begin{aligned}\mathbf{0} = \sum_{t=1}^{T}\Big(&\boldsymbol{y}_t(\boldsymbol{y}_t)^\top - \boldsymbol{y}_t(\widetilde{\mathbf{x}}_t)^\top\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{x}}_t(\boldsymbol{y}_t)^\top - \boldsymbol{y}_t\mathbf{a}^\top - \mathbf{a}(\boldsymbol{y}_t)^\top \\ &+ \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}(\widetilde{\mathbf{x}}_t)^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big) - T\mathbf{R}\end{aligned} \tag{46}$$

We can then solve for $\mathbf{R}$, giving us the new $\mathbf{R}$ that maximizes Ψ,

$$\begin{aligned}\mathbf{R}_{\text{new}} &= \frac{1}{T}\sum_{t=1}^{T}\Big(\boldsymbol{y}_t(\boldsymbol{y}_t)^\top - \boldsymbol{y}_t(\widetilde{\mathbf{x}}_t)^\top\mathbf{Z}^\top - \mathbf{Z}\widetilde{\mathbf{x}}_t(\boldsymbol{y}_t)^\top - \boldsymbol{y}_t\mathbf{a}^\top - \mathbf{a}(\boldsymbol{y}_t)^\top \\ &\quad + \mathbf{Z}\widetilde{\mathbf{P}}_t\mathbf{Z}^\top + \mathbf{Z}\widetilde{\mathbf{x}}_t\mathbf{a}^\top + \mathbf{a}(\widetilde{\mathbf{x}}_t)^\top\mathbf{Z}^\top + \mathbf{a}\mathbf{a}^\top\Big) \\ &= \frac{1}{T}\sum_{t=1}^{T}\Big((\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})^\top + \mathbf{Z}(\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{x}}_t(\widetilde{\mathbf{x}}_t)^\top)\mathbf{Z}^\top\Big) \\ &= \frac{1}{T}\sum_{t=1}^{T}\Big((\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})^\top + \mathbf{Z}\widetilde{\mathbf{V}}_t\mathbf{Z}^\top\Big)\end{aligned} \tag{47}$$

As with $\mathbf{Q}$, this derivation immediately generalizes to a block diagonal matrix:

$$\mathbf{R} = \begin{bmatrix}\mathbf{R}_1 & 0 & 0 \\ 0 & \mathbf{R}_2 & 0 \\ 0 & 0 & \mathbf{R}_3\end{bmatrix}$$

In this case,

$$\mathbf{R}_{i,\text{new}} = \frac{1}{T}\sum_{t=1}^{T}\Big((\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})^\top + \mathbf{Z}\widetilde{\mathbf{V}}_t\mathbf{Z}^\top\Big)_i \tag{48}$$

where the subscript i means we take the elements in the matrix in the big parentheses that are analogous to $\mathbf{R}_i$. If $\mathbf{R}_i$ is comprised of rows a to b and columns c to d of matrix $\mathbf{R}$, then we take rows a to b and columns c to d of matrix subscripted by i in equation (48).

Dealing with missing values in the data is straight-forward if $\mathbf{R}$ is constrained to be strictly diagonal (Shumway and Stoffer, 2006). If $\mathbf{R}$ is diagonal (not block diagonal) and the diagonal elements are unequal (or at least not forced to be shared), then the update equation for $\mathbf{R}$ becomes

$$\begin{aligned}\text{diag}(\mathbf{R}_{\text{new}}) &= \frac{1}{T}\sum_{t=1}^{T}\text{diag}\Big((\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})^\top + \mathbf{Z}\widetilde{\mathbf{V}}_t\mathbf{Z}^\top\Big) \\ &= \frac{1}{T}\sum_{t=1}^{T}\text{diag}(\mathbf{J}_t)\end{aligned} \tag{49}$$

where diag means "the diagonal of". If the i-th value of $\boldsymbol{y}$ is missing at time t, $\boldsymbol{y}_{i,t}$, then the (i,i) value of $\mathbf{R}$ from the previous iteration of the EM algorithm, $\mathbf{R}_{i,i,\text{old}}$, is used in place of the t-th (i,i) value of matrix $\mathbf{J}_t$ in the summation. See Shumway and Stoffer (2006) for a discussion of the $\mathbf{R}$ update equations when $\mathbf{R}$ has non-diagonal elements and there are missing values.

3.7 Update equation for $\boldsymbol{\xi}$ (unconstrained)

Take the derivative of Ψ with respect to $\boldsymbol{\xi}$. Terms not involving $\boldsymbol{\xi}$, equal 0 and drop out.

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\xi} = -\frac{1}{2}\big(&- \mathrm{E}[\partial((\boldsymbol{\xi})^\top(\mathbf{V}_1)^{-1}\boldsymbol{x}_1)/\partial\boldsymbol{\xi}] - \mathrm{E}[\partial((\boldsymbol{x}_1)^\top(\mathbf{V}_1)^{-1}\boldsymbol{\xi})/\partial\boldsymbol{\xi}] \\ &+ \partial(\boldsymbol{\xi}^\top(\mathbf{V}_1)^{-1}\boldsymbol{\xi})/\partial\boldsymbol{\xi}\big)\end{aligned} \tag{50}$$

Using relations (6) and (10) and using $(\mathbf{V}_1)^{-1} = ((\mathbf{V}_1)^{-1})^\top$, we have

$$\partial\Psi/\partial\boldsymbol{\xi} = -\frac{1}{2}\big(- \mathrm{E}[(\boldsymbol{x}_1)^\top(\mathbf{V}_1)^{-1}] - \mathrm{E}[(\boldsymbol{x}_1)^\top(\mathbf{V}_1)^{-1}] + 2\boldsymbol{\xi}^\top(\mathbf{V}_1)^{-1}\big) \tag{51}$$

Pulling the parameters out of the expectations, we get

$$\partial\Psi/\partial\boldsymbol{\xi} = -\frac{1}{2}\big(- 2\,\mathrm{E}[(\boldsymbol{x}_1)^\top](\mathbf{V}_1)^{-1} + 2\boldsymbol{\xi}^\top(\mathbf{V}_1)^{-1}\big) \tag{52}$$

We then set the left side to zero, take the transpose, and cancel out $-1/2$ and $(\mathbf{V}_1)^{-1}$ (by noting that it is a variance-covariance matrix and is invertable).

$$\mathbf{0} = \big((\mathbf{V}_1)^{-1}\,\mathrm{E}[\boldsymbol{x}_1] + (\mathbf{V}_1)^{-1}\boldsymbol{\xi}\big) = (\widetilde{\mathbf{x}}_1 - \boldsymbol{\xi}) \tag{53}$$

Thus,

$$\boldsymbol{\xi}_{\mathrm{new}} = \widetilde{\mathbf{x}}_1 \tag{54}$$

4 The constrained update equations

The previous sections dealt with the case where all the elements in a parameter matrix are estimated. In this section, I deal with the case where some of the elements are constrained, for example when some elements are fixed values and some elements are shared (meaning they are forced to have the same value). One cannot simply use the elements from the unconstrained case for the free elements because the solution depends on the fixed values; those have to be included in the solution. One could always go through each matrix element one-by-one, but that would be very slow since the Kalman smoother would need to be run after updating each matrix element. Rather one would like to find a simultaneous solution for all the free elements in our constrained parameter matrix.

Let's say we have some parameter matrix $\mathbf{M}$ (here $\mathbf{M}$ could be any of the parameters in the MARSS model) with fixed, shared and unshared elements:

$$\mathbf{M} = \begin{bmatrix} a & 0.9 & c \\ -1.2 & a & 0 \\ 0 & c & b \end{bmatrix}$$

The matrix $\mathbf{M}$ can be rewritten in terms of a fixed and free part, where in the fixed part all free elements are set to zero and in the free part all fixed elements

are set to zero:

$$\mathbf{M} = \begin{bmatrix} 0 & 0.9 & 0 \\ -1.2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} a & 0 & c \\ 0 & a & 0 \\ 0 & c & b \end{bmatrix} = \mathbf{M}_{\text{fixed}} + \mathbf{M}_{\text{free}}$$

The vec function turns any matrix into a column vector by stacking the columns on top of each other. Thus,

$$\text{vec}(\mathbf{M}) = \begin{bmatrix} a \\ -1.2 \\ 0 \\ 0.9 \\ a \\ c \\ c \\ 0 \\ b \end{bmatrix}$$

We can now write $\text{vec}(\mathbf{M})$ as a linear combination of $\mathbf{f} = \text{vec}(\mathbf{M}_{\text{fixed}})$ and $\mathbf{Dm} = \text{vec}(\mathbf{M}_{\text{free}})$. $\mathbf{m}$ is a $p \times 1$ column vector of the p free values, in this case $p = 3$ and the free values are a, b, c. $\mathbf{D}$ is a design matrix that translates $\mathbf{m}$ into $\text{vec}(\mathbf{M}_{\text{free}})$. For example,

$$\text{vec}(\mathbf{M}) = \begin{bmatrix} a \\ -1.2 \\ 0 \\ 0.9 \\ a \\ c \\ c \\ 0 \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ -1.2 \\ 0 \\ 0.9 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \mathbf{f} + \mathbf{Dm}$$

The derivation proceeds by rewriting the likelihood as a function of $\text{vec}(\mathbf{M})$, where $\mathbf{M}$ is whatever parameter matrix for which one is deriving the update equation. Then one rewrites that as a function of $\mathbf{m}$ using the relationship $\text{vec}(\mathbf{M}) = \mathbf{f} + \mathbf{Dm}$. Finally, one finds the $\mathbf{m}$ that sets the derivative of Ψ with respect to $\mathbf{m}$ to zero. Conceptually, the algebraic steps in the derivation are similar to those in the unconstrained derivation. Thus, I will leave out most of the intermediate steps. The derivations require a few new matrix algebra and vec relationships shown in Table 2.

4.1 The general u update equations

Since $\mathbf{u}$ is already a column vector, it can be rewritten simply as $\mathbf{u} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{v}$, where $\boldsymbol{v}$ is the column vector of estimated parameters in $\mathbf{u}$. We then solve for $\partial\Psi/\partial\boldsymbol{v}$ by replacing $\mathbf{u}$ with $\mathbf{u} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{v}$ in the expected log likelihood

Table 2: Kronecker and vec relations. Here $\mathbf{A}$ is $n \times m$, $\mathbf{B}$ is $m \times p$, $\mathbf{C}$ is $p \times q$. $\mathbf{a}$ is a $m \times 1$ column vector and $\mathbf{b}$ is a $p \times 1$ column vector. The symbol $\otimes$ stands for the Kronecker product: $\mathbf{A} \otimes \mathbf{C}$ is a $np \times mq$ matrix. The identity matrix, $\mathbf{I}_n$, is a $n \times n$ diagonal matrix with ones on the diagonal.

$$\text{vec}(\mathbf{a}) = \text{vec}(\mathbf{a}^\top) = \mathbf{a}$$
The vec of a column vector (or its transpose) is itself. (55)

$$\text{vec}(\mathbf{Aa}) = (\mathbf{a}^\top \otimes \mathbf{I}_n)\,\text{vec}(\mathbf{A}) = \mathbf{Aa}$$
$\text{vec}(\mathbf{Aa}) = \mathbf{Aa}$ since $\mathbf{Aa}$ is itself an $m \times 1$ column vector. (56)

$$\text{vec}(\mathbf{AB}) = (\mathbf{I}_p \otimes \mathbf{A})\,\text{vec}(\mathbf{B}) = (\mathbf{B}^\top \otimes \mathbf{I}_n)\,\text{vec}(\mathbf{A}) \tag{57}$$

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\,\text{vec}(\mathbf{B}) \tag{58}$$

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC} \otimes \mathbf{BD}) \tag{59}$$

$$(\mathbf{a} \otimes \mathbf{I}_p)\mathbf{C} = (\mathbf{a} \otimes \mathbf{C})$$
$$\mathbf{C}(\mathbf{a}^\top \otimes \mathbf{I}_q) = (\mathbf{a}^\top \otimes \mathbf{C}) \tag{60}$$

$$(\mathbf{a} \otimes \mathbf{I}_p)\mathbf{C}(\mathbf{b}^\top \otimes \mathbf{I}_q) = (\mathbf{ab}^\top \otimes \mathbf{C}) \tag{61}$$

$$(\mathbf{a} \otimes \mathbf{a}) = \text{vec}(\mathbf{aa}^\top)$$
$$(\mathbf{a}^\top \otimes \mathbf{a}^\top) = (\mathbf{a} \otimes \mathbf{a})^\top = (\,\text{vec}(\mathbf{aa}^\top))^\top \tag{62}$$

---

function. In the derivation below, the u subscripts on $\mathbf{f}$ and $\mathbf{D}$ have been left off to remove clutter.

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{v} = -\frac{1}{2}\sum_{t=2}^{T}\Big(&- \mathrm{E}[\partial((\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}(\mathbf{f}+\mathbf{D}\boldsymbol{v}))/\partial\boldsymbol{v}] \\ &- \mathrm{E}[\partial((\mathbf{f}+\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\boldsymbol{x}_t)/\partial\boldsymbol{v}] + \mathrm{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top\mathbf{Q}^{-1}(\mathbf{f}+\mathbf{D}\boldsymbol{v}))/\partial\boldsymbol{v}] \\ &+ \mathrm{E}[\partial((\mathbf{f}+\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1})/\partial\boldsymbol{v}] + \partial((\mathbf{f}+\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}(\mathbf{f}+\mathbf{D}\boldsymbol{v}))/\partial\boldsymbol{v}\Big)\end{aligned} \tag{63}$$

The terms involving only $\mathbf{f}$ drop out (because they don't involve $\boldsymbol{v}$). This gives

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{v} = -\frac{1}{2}\sum_{t=2}^{T}\Big(&- \mathrm{E}[\partial((\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v})/\partial\boldsymbol{v}] - \mathrm{E}[\partial((\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\boldsymbol{x}_t)/\partial\boldsymbol{v}] \\ &+ \mathrm{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v})/\partial\boldsymbol{v}] + \mathrm{E}[\partial((\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1})/\partial\boldsymbol{v}] \\ &+ \partial(\mathbf{f}^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v})/\partial\boldsymbol{v} + \partial((\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\mathbf{f})/\partial\boldsymbol{v} + \partial((\mathbf{D}\boldsymbol{v})^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v})/\partial\boldsymbol{v}\Big)\end{aligned} \tag{64}$$

Using the matrix differentiation relations in Table 1, we get

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{v} = -\frac{1}{2}\sum_{t=2}^{T}\Big(&- 2\,\mathrm{E}[(\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}\mathbf{D}] + 2\,\mathrm{E}[(\mathbf{B}\boldsymbol{x}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{D}] \\ &+ 2\mathbf{f}^\top\mathbf{Q}^{-1}\mathbf{D} + 2\boldsymbol{v}^\top\mathbf{D}^\top\mathbf{Q}^{-1}\mathbf{D}\Big)\end{aligned} \tag{65}$$

Set the left side to zero and transpose the whole equation. Then we solve for $\boldsymbol{v}$.

$$\mathbf{0} = \sum_{t=2}^{T}\Big(\mathbf{D}^\top\mathbf{Q}^{-1}(\,\mathrm{E}[\boldsymbol{x}_t] - \mathbf{B}\,\mathrm{E}[\boldsymbol{x}_{t-1}] - \mathbf{f}) - \mathbf{D}^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v}\Big) \tag{66}$$

Thus,

$$(T-1)\mathbf{D}^\top\mathbf{Q}^{-1}\mathbf{D}\boldsymbol{v} = \mathbf{D}^\top\mathbf{Q}^{-1}\sum_{t=2}^{T}\big(\,\mathrm{E}[\boldsymbol{x}_t] - \mathbf{B}\,\mathrm{E}[\boldsymbol{x}_{t-1}] - \mathbf{f}\big) \tag{67}$$

Thus, the updated $\boldsymbol{v}$ is

$$\boldsymbol{v}_{\mathrm{new}} = \frac{1}{T-1}\big(\mathbf{D}^\top\mathbf{Q}^{-1}\mathbf{D}\big)^{-1}\mathbf{D}^\top\mathbf{Q}^{-1}\sum_{t=2}^{T}\big(\widetilde{\mathbf{x}}_t - \mathbf{B}\widetilde{\mathbf{x}}_{t-1} - \mathbf{f}\big) \tag{68}$$

and

$$\mathbf{u}_{\mathrm{new}} = \mathbf{f}_u + \mathbf{D}_u\boldsymbol{v}_{\mathrm{new}} \tag{69}$$

If $\mathbf{Q}$ is diagonal or the fixed values are all 0, this will reduce just updating the free elements in $\mathbf{u}$ using their values from the unconstrained update equation.

4.2 The general a update equation

The derivation of the update equation for $\mathbf{a}$ with fixed and shared values is completely analogous to the derivation for $\mathbf{u}$. If $\mathbf{a} = \mathbf{f}_a + \mathbf{D}_a\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is a column vector of the estimated values then

$$\boldsymbol{\alpha}_{\text{new}} = \frac{1}{T}\left(\mathbf{D}_a^\top \mathbf{R}^{-1}\mathbf{D}_a\right)^{-1}\mathbf{D}_a^\top \mathbf{R}^{-1}\sum_{t=1}^{T}\left(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{f}_a\right) \tag{70}$$

The new $\mathbf{a}$ parameter is then

$$\mathbf{a}_{\text{new}} = \mathbf{f}_a + \mathbf{D}_a\boldsymbol{\alpha}_{\text{new}} \tag{71}$$

Again if $\mathbf{R}$ is diagonal or the fixed values are all 0, this will reduce just updating the free elements in $\mathbf{a}$ using their values from the unconstrained update equation. The modification for missing values follows the unconstrained case. Specifically, when $y_{i,t}$ is missing, the $\mathbf{a}_{i\text{old}}$ value is used for the i-th value of $(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{f}_a)$ at time t.

4.3 The general $\boldsymbol{\xi}$ update equation

The derivation of the update equation for $\boldsymbol{\xi}$ with fixed and shared values is similar to the derivation for $\mathbf{u}$ and $\mathbf{a}$. If $\boldsymbol{\xi} = \mathbf{f}_\xi + \mathbf{D}_\xi\mathbf{p}$, where $\mathbf{p}$ is a column vector of the estimated values then

$$\partial\Psi/\partial\mathbf{p} = \left((\widetilde{\mathbf{x}}_1)^\top(\mathbf{V}_1)^{-1} - \boldsymbol{\xi}^\top(\mathbf{V}_1)^{-1}\right)\mathbf{D} \tag{72}$$

Replace $\boldsymbol{\xi}$ with $\mathbf{f}_\xi + \mathbf{D}_\xi\mathbf{p}$, set the left side to zero and transpose:

$$\mathbf{0} = \mathbf{D}^\top\left((\mathbf{V}_1)^{-1}\widetilde{\mathbf{x}}_1 - (\mathbf{V}_1)^{-1}\mathbf{f}_\pi + (\mathbf{V}_1)^{-1}\mathbf{D}\mathbf{p}\right) \tag{73}$$

Thus,

$$\mathbf{p}_{\text{new}} = \left(\mathbf{D}_\xi^\top(\mathbf{V}_1)^{-1}\mathbf{D}_\xi\right)^{-1}\mathbf{D}_\xi^\top(\mathbf{V}_1)^{-1}(\widetilde{\mathbf{x}}_1 - \mathbf{f}_\xi) \tag{74}$$

The new $\boldsymbol{\xi}$ is then,

$$\boldsymbol{\xi}_{\text{new}} = \mathbf{f}_\xi + \mathbf{D}_\xi\mathbf{p}_{\text{new}} \tag{75}$$

4.4 The general B update equation

The matrix $\mathbf{B}$ is rewritten as $\mathbf{B} = \mathbf{B}_{\text{fixed}} + \mathbf{B}_{\text{free}}$, thus $\text{vec}(\mathbf{B}) = \mathbf{f}_b + \mathbf{D}_b\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the $p \times 1$ column vector of the p estimated values, $\mathbf{f}_b = \text{vec}(\mathbf{B}_{\text{fixed}})$ and $\mathbf{D}_b\boldsymbol{\beta} = \text{vec}(\mathbf{B}_{\text{free}})$. Take the derivative of Ψ with respect to $\boldsymbol{\beta}$; terms in Ψ that do not involve $\mathbf{B}$ also do not involve $\boldsymbol{\beta}$ so they will equal 0 and drop out.

$$\begin{aligned}\partial\Psi/\partial\boldsymbol{\beta} = -\frac{1}{2}\sum_{t=2}^{T}\Big(&-\text{E}[\partial((\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1})/\partial\boldsymbol{\beta}]\\ &-\text{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top\mathbf{Q}^{-1}\boldsymbol{x}_t)/\partial\boldsymbol{\beta}] + \text{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top\mathbf{Q}^{-1}(\mathbf{B}\boldsymbol{x}_{t-1}))/\partial\boldsymbol{\beta}]\\ &+\text{E}[\partial((\mathbf{B}\boldsymbol{x}_{t-1})^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\beta}] + \text{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1})/\partial\boldsymbol{\beta}]\Big)\end{aligned} \tag{76}$$

This needs to be rewritten as a function of $\boldsymbol{\beta}$ instead of $\mathbf{B}$. Note that $\mathbf{B}\boldsymbol{x}_{t-1}$ is a column vector and use relation (56) to show that:

$$\begin{aligned} \mathbf{B}\boldsymbol{x}_{t-1} = \operatorname{vec}(\mathbf{B}\boldsymbol{x}_{t-1}) = \mathbf{K}_b \operatorname{vec}(\mathbf{B}) = \mathbf{K}_b(\mathbf{f}_b + \mathbf{D}_b\boldsymbol{\beta}), \\ \text{where } \mathbf{K}_b = ((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I}) \end{aligned} \tag{77}$$

Thus, $\partial\Psi/\partial\boldsymbol{\beta}$ becomes (the b subscripts are left off $\mathbf{K}$, $\mathbf{F}$ and $\mathbf{D}$ to remove clutter):

$$\begin{aligned} \partial\Psi/\partial\boldsymbol{\beta} = -\frac{1}{2}\sum_{t=2}^{T}\Big(& - \operatorname{E}[\partial((\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))/\partial\boldsymbol{\beta}] \\ & - \operatorname{E}[\partial((\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))^\top\mathbf{Q}^{-1}\boldsymbol{x}_t)/\partial\boldsymbol{\beta}] \\ & + \operatorname{E}[\partial((\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))^\top\mathbf{Q}^{-1}(\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta})))/\partial\boldsymbol{\beta}] \\ & + \operatorname{E}[\partial((\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))^\top\mathbf{Q}^{-1}\mathbf{u})/\partial\boldsymbol{\beta}] + \operatorname{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{K}(\mathbf{f}+\mathbf{D}\boldsymbol{\beta}))/\partial\boldsymbol{\beta}]\Big) \end{aligned} \tag{78}$$

After a bit of matrix algebra and remembering that $\partial(\mathbf{a}^\top\mathbf{c})/\partial\mathbf{a} = \partial(\mathbf{c}^\top\mathbf{a})/\partial\mathbf{a}$, equation (6), and that partial derivatives of constants equal 0, the above can be simplified to

$$\begin{aligned} \partial\Psi/\partial\boldsymbol{\beta} = -\frac{1}{2}\sum_{t=2}^{T}\Big(& - 2\operatorname{E}[\partial((\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}\boldsymbol{\beta})/\partial\boldsymbol{\beta}] \\ & + 2\operatorname{E}[\partial((\mathbf{K}\mathbf{f})^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}\boldsymbol{\beta})/\partial\boldsymbol{\beta}] \\ & + \operatorname{E}[\partial(\boldsymbol{\beta}^\top(\mathbf{K}\mathbf{D})^\top\mathbf{Q}^{-1}(\mathbf{K}\mathbf{D})\boldsymbol{\beta})/\partial\boldsymbol{\beta}] + 2\operatorname{E}[\partial(\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}\boldsymbol{\beta})/\partial\boldsymbol{\beta}]\Big) \end{aligned} \tag{79}$$

Using relations (6) and (10), using $\mathbf{Q}^{-1} = (\mathbf{Q}^{-1})^\top$, and getting rid of the $-1/2$, we have

$$\begin{aligned} \partial\Psi/\partial\boldsymbol{\beta} = \sum_{t=2}^{T}\Big(& \operatorname{E}[(\boldsymbol{x}_t)^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}] - \operatorname{E}[(\mathbf{K}\mathbf{f})^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}] \\ & + \operatorname{E}[\boldsymbol{\beta}^\top(\mathbf{K}\mathbf{D})^\top\mathbf{Q}^{-1}(\mathbf{K}\mathbf{D})] - \operatorname{E}[\mathbf{u}^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{D}]\Big) \end{aligned} \tag{80}$$

The left side can be set to 0 (a $1 \times p$ matrix) and the whole equation transposed, giving:

$$\begin{aligned} \mathbf{0} = \sum_{t=2}^{T}\Big(& \operatorname{E}[(\mathbf{K}\mathbf{D})^\top\mathbf{Q}^{-1}\boldsymbol{x}_t] - \operatorname{E}[(\mathbf{K}\mathbf{D})^\top\mathbf{Q}^{-1}\mathbf{K}\mathbf{f}] \\ & + \operatorname{E}[(\mathbf{K}\mathbf{D})^\top\mathbf{Q}^{-1}(\mathbf{K}\mathbf{D})]\boldsymbol{\beta} - \operatorname{E}[(\mathbf{K}\mathbf{D})^\top\mathbf{Q}^{-1}\mathbf{u}]\Big) \end{aligned} \tag{81}$$

Replacing $\mathbf{K}$ with $((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I})$, we have

$$\begin{aligned}\mathbf{0} = \sum_{t=2}^{T} \Big(& \mathrm{E}[(((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I})\mathbf{D})^\top \mathbf{Q}^{-1}\boldsymbol{x}_t] \\ & - \mathrm{E}[(((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I})\mathbf{D})^\top \mathbf{Q}^{-1}((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I})\mathbf{f}] \\ & + \mathrm{E}[(((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I})\mathbf{D})^\top \mathbf{Q}^{-1}((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I})\mathbf{D}]\boldsymbol{\beta} \\ & - \mathrm{E}[(((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I})\mathbf{D})^\top \mathbf{Q}^{-1}\mathbf{u}]\Big)\end{aligned} \tag{82}$$

This looks daunting, but using relation (56) and noting that $(\mathbf{A}\otimes\mathbf{B})^\top = (\mathbf{A}^\top \otimes \mathbf{B}^\top)$, we can simplify equation (82) using the following:

$$\begin{aligned}&((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I})\mathbf{Q}^{-1}\mathbf{u} = (\boldsymbol{x}_{t-1} \otimes \mathbf{I})\mathbf{Q}^{-1}\mathbf{u} \\ &\quad = (\boldsymbol{x}_{t-1} \otimes \mathbf{I})\,\mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}),\ \text{because}\ \mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u})\text{is a column vector} \\ &\quad = \mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}(\boldsymbol{x}_{t-1})^\top), \text{using relation (56)}\end{aligned}$$

Similarly,

$$((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I})\mathbf{Q}^{-1}\boldsymbol{x}_t = \mathrm{vec}(\mathbf{Q}^{-1}\boldsymbol{x}_t(\boldsymbol{x}_{t-1})^\top)$$

Using relation (61):

$$(\boldsymbol{x}_{t-1} \otimes \mathbf{I}_m)^\top \mathbf{Q}^{-1}((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I}_m)\mathbf{f} = (\boldsymbol{x}_{t-1}(\boldsymbol{x}_{t-1})^\top \otimes \mathbf{Q}^{-1})\mathbf{f}$$

Similarly,

$$(\boldsymbol{x}_{t-1} \otimes \mathbf{I})^\top \mathbf{Q}^{-1}((\boldsymbol{x}_{t-1})^\top \otimes \mathbf{I})\mathbf{D}\boldsymbol{\beta} = (\boldsymbol{x}_{t-1}(\boldsymbol{x}_{t-1})^\top \otimes \mathbf{Q}^{-1})\mathbf{D}\boldsymbol{\beta}$$

Using these simplifications in equation (82), we get

$$\begin{aligned}\mathbf{0} = \sum_{t=2}^{T} \Big(& \mathrm{E}[\mathbf{D}^\top \mathrm{vec}(\mathbf{Q}^{-1}\boldsymbol{x}_t(\boldsymbol{x}_{t-1})^\top)] - \mathrm{E}[\mathbf{D}^\top(\boldsymbol{x}_{t-1}(\boldsymbol{x}_{t-1})^\top \otimes \mathbf{Q}^{-1})\mathbf{f}] \\ & - \mathrm{E}[\mathbf{D}^\top(\boldsymbol{x}_{t-1}(\boldsymbol{x}_{t-1})^\top \otimes \mathbf{Q}^{-1})\mathbf{D}]\boldsymbol{\beta} - \mathrm{E}[\mathbf{D}^\top \mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}(\boldsymbol{x}_{t-1})^\top)]\Big)\end{aligned} \tag{83}$$

Replacing the expectations with the Kalman smoother output, we arrive at:

$$\begin{aligned}\mathbf{0} = \sum_{t=2}^{T} \Big(& \mathbf{D}^\top \mathrm{vec}(\mathbf{Q}^{-1}\widetilde{\mathbf{P}}_{t,t-1}) - \mathbf{D}^\top(\widetilde{\mathbf{P}}_{t-1} \otimes \mathbf{Q}^{-1})\mathbf{f} \\ & - \mathbf{D}^\top(\widetilde{\mathbf{P}}_{t-1} \otimes \mathbf{Q}^{-1})\mathbf{D}\boldsymbol{\beta} - \mathbf{D}^\top \mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top)\Big)\end{aligned} \tag{84}$$

Solving for $\boldsymbol{\beta}$,

$$\begin{aligned}\boldsymbol{\beta}_{\text{new}} = & \Big(\sum_{t=2}^{T} \mathbf{D}^\top(\widetilde{\mathbf{P}}_{t-1} \otimes \mathbf{Q}^{-1})\mathbf{D} \Big)^{-1} \mathbf{D}^\top \Big(\sum_{t=2}^{T} \big(\mathrm{vec}(\mathbf{Q}^{-1}\widetilde{\mathbf{P}}_{t,t-1}) \\ & - (\widetilde{\mathbf{P}}_{t-1} \otimes \mathbf{Q}^{-1})\mathbf{f} - \mathrm{vec}(\mathbf{Q}^{-1}\mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top)\big)\Big)\end{aligned} \tag{85}$$

This requires that $(\mathbf{D}^\top(\widetilde{\mathbf{P}}_{t-1} \otimes \mathbf{Q}^{-1})\mathbf{D})$ is invertable, which it is because it is a $p \times p$ diagonal matrix with only non-zero values on the diagonal.

Combining $\boldsymbol{\beta}_{\text{new}}$ with $\mathbf{B}_{\text{fixed}}$, we arrive at the vec of the updated $\mathbf{B}$ matrix:

$$\text{vec}(\mathbf{B}_{\text{new}}) = \mathbf{f}_b + \mathbf{D}_b\boldsymbol{\beta}_{\text{new}} \tag{86}$$

When there are no fixed or shared values, $\mathbf{B}_{\text{fixed}}$ equals zero and $\mathbf{D}_b$ equals an identity matrix. Equation (85) then reduces to the unconstrained form. To see this take the vec of the unconstrained update equation for $\mathbf{B}$ and notice that $\mathbf{Q}^{-1}$ can be then factored out.

4.5 The general Z update equation

The derivation of the update equation for $\mathbf{Z}$ with fixed and shared values is analogous to the derivation for $\mathbf{B}$. The matrix $\mathbf{Z}$ is rewritten as $\mathbf{Z} = \mathbf{Z}_{\text{fixed}} + \mathbf{Z}_{\text{free}}$, thus $\text{vec}(\mathbf{Z}) = \mathbf{f}_z + \mathbf{D}_z\boldsymbol{\zeta}$, where $\boldsymbol{\zeta}$ is the column vector of the p estimated values, $\mathbf{f}_z = \text{vec}(\mathbf{Z}_{\text{fixed}})$ and $\mathbf{D}_z\boldsymbol{\zeta} = \text{vec}(\mathbf{Z}_{\text{free}})$.

$$\begin{aligned}\boldsymbol{\zeta}_{\text{new}} = \Bigg(\sum_{t=1}^{T}(\mathbf{D}_z^\top(\widetilde{\mathbf{P}}_t \otimes \mathbf{R}^{-1})\mathbf{D}_z)\Bigg)^{-1}\mathbf{D}_z^\top\Bigg(\sum_{t=1}^{T}\big(\text{vec}(\mathbf{R}^{-1}\boldsymbol{y}_t(\widetilde{\mathbf{x}}_t)^\top) \\ -(\widetilde{\mathbf{P}}_t \otimes \mathbf{R}^{-1})\mathbf{f} - \text{vec}(\mathbf{R}^{-1}\mathbf{a}(\widetilde{\mathbf{x}}_t)^\top)\big)\Bigg)\end{aligned} \tag{87}$$

Combining $\boldsymbol{\zeta}_{\text{new}}$ with $\mathbf{Z}_{\text{fixed}}$, we arrive at the vec of the updated $\mathbf{Z}$ matrix:

$$\text{vec}(\mathbf{Z}_{\text{new}}) = \mathbf{f}_z + \mathbf{D}_z\boldsymbol{\zeta}_{\text{new}} \tag{88}$$

4.6 The general Q update equation

A general analytical solution for fixed and shared elements in $\mathbf{Q}$ is problematic since the inverse of $\mathbf{Q}$ appears in the likelihood and since $\mathbf{Q}^{-1}$ cannot always be rewritten as a function of $\text{vec}(\mathbf{Q})$. It might be an option to use numerical maximization of $\partial\Psi/\partial q_{i,j}$ where $q_{i,j}$ is a free element in $\mathbf{Q}$, but this will slow down the algorithm enormously. However, in a few important special – yet quite broad — cases, an analytical solution can be derived. The most general of these special cases is a block-symmetric matrix with optional independent fixed blocks (subsection 4.6.5). Indeed, all other cases (diagonal, block-diagonal, unconstrained, equal variance-covariance) except one (a replicated block-diagonal) are special cases of the blocked matrix with optional independent fixed blocks.

The general update equation is

$$\begin{aligned}\boldsymbol{q}_{\text{new}} = \frac{1}{T-1}(\mathbf{D}_q^\top\mathbf{D}_q)^{-1}\mathbf{D}_q^\top\,\text{vec}(\mathbf{S}) \\ \text{vec}(\mathbf{Q})_{\text{new}} = \mathbf{f}_q + \mathbf{D}_q\boldsymbol{q}_{\text{new}}\end{aligned} \tag{89}$$

where $\mathbf{f}_q$, $\mathbf{D}_q$, and $\boldsymbol{q}$ have their standard definitions (section 4). The vec of $\mathbf{Q}$ is written in the form of $\text{vec}(\mathbf{Q}) = \mathbf{f}_q + \mathbf{D}_q\boldsymbol{q}$, where $\mathbf{f}_q$ is a $m^2 \times 1$ column vector

of the fixed values including zero, $\mathbf{D}_q$ is the $m^2 \times p$ design matrix, and $\boldsymbol{q}$ is a column vector of the p free values.

Below I show how the update equation arises by working through a few of the special cases.

4.6.1 Special case: diagonal Q matrix (with shared or unique parameters)

Let $\mathbf{Q}$ be some diagonal matrix with fixed and shared values. For example,

$$\mathbf{Q} = \begin{bmatrix} q_1 & 0 & 0 & 0 & 0 \\ 0 & f_1 & 0 & 0 & 0 \\ 0 & 0 & q_2 & 0 & 0 \\ 0 & 0 & 0 & f_2 & 0 \\ 0 & 0 & 0 & 0 & q_2 \end{bmatrix}$$

Here, f's are fixed values (constants) and q's are free parameters elements. The vec of $\mathbf{Q}^{-1}$ can be written then as $\text{vec}(\mathbf{Q}^{-1}) = \mathbf{f}_q^* + \mathbf{D}_q \boldsymbol{q}^*$, where $\mathbf{f}^*$ is like $\mathbf{f}_q$ but with the corresponding i-th non-zero fixed values replaced by $1/f_i$ and $\boldsymbol{q}^*$ is a column vector of 1 over the q_i values. For our example above,

$$\boldsymbol{q}^* = \begin{bmatrix} 1/q_1 \\ 1/q_2 \end{bmatrix}$$

Take the partial derivative of Ψ with respect to $\boldsymbol{q}^*$. We can do this because $\mathbf{Q}^{-1}$ is diagonal and thus each element of $\boldsymbol{q}^*$ is independent of the other elements; otherwise we would not necessarily be able to vary one element of $\boldsymbol{q}^*$ while holding the other elements constant.

$$\begin{aligned} \partial\Psi/\partial\boldsymbol{q}^* = -\frac{1}{2}\sum_{t=2}^{T} \partial\Big(& \text{E}[(\boldsymbol{x}_t)^\top \mathbf{Q}^{-1}\boldsymbol{x}_t] - \text{E}[(\boldsymbol{x}_t)^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1}] \\ & - \text{E}[(\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1}\boldsymbol{x}_t] - \text{E}[(\boldsymbol{x}_t)^\top \mathbf{Q}^{-1}\mathbf{u}] \\ & - \text{E}[\mathbf{u}^\top \mathbf{Q}^{-1}\boldsymbol{x}_t] + \text{E}[(\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1}] \\ & + \text{E}[(\mathbf{B}\boldsymbol{x}_{t-1})^\top \mathbf{Q}^{-1}\mathbf{u}] + \text{E}[\mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{B}\boldsymbol{x}_{t-1}] + \mathbf{u}^\top \mathbf{Q}^{-1}\mathbf{u}\Big)/\partial\boldsymbol{q}^* \\ & - \partial\big(\frac{T-1}{2}\log|\mathbf{Q}|\big)/\partial\boldsymbol{q}^* \end{aligned} \tag{90}$$

Using the same vec operations as in the derivations for $\mathbf{B}$ and $\mathbf{Z}$, pull $\mathbf{Q}^{-1}$ out from the middle and replace the expectations with the Kalman smoother

output.[8]

$$
\begin{aligned}
\partial\Psi/\partial\boldsymbol{q}^{*} = -\frac{1}{2}\sum_{t=2}^{T}\partial\Big(&\mathrm{E}[((\boldsymbol{x}_t)^\top\otimes(\boldsymbol{x}_t)^\top)] - \mathrm{E}[((\boldsymbol{x}_t)^\top\otimes(\mathbf{B}\boldsymbol{x}_{t-1})^\top)]\\
&- \mathrm{E}[((\mathbf{B}\boldsymbol{x}_{t-1})^\top\otimes(\boldsymbol{x}_t)^\top)] - \mathrm{E}[((\boldsymbol{x}_t)^\top\otimes(\mathbf{u})^\top)]\\
&- \mathrm{E}[(\mathbf{u}^\top\otimes(\boldsymbol{x}_t)^\top)] + \mathrm{E}[((\mathbf{B}\boldsymbol{x}_{t-1})^\top\otimes(\mathbf{B}\boldsymbol{x}_{t-1})^\top)]\\
&+ \mathrm{E}[((\mathbf{B}\boldsymbol{x}_{t-1})^\top\otimes(\mathbf{u})^\top)] + \mathrm{E}[(\mathbf{u}^\top\otimes(\mathbf{B}\boldsymbol{x}_{t-1})^\top)] + (\mathbf{u}^\top\otimes\mathbf{u}^\top)\Big)\,\mathrm{vec}(\mathbf{Q}^{-1})/\partial\boldsymbol{q}^{*}\\
-\partial\big(\frac{T-1}{2}\log|\mathbf{Q}|\big)/\partial\boldsymbol{q}^{*}&\\
= -\frac{1}{2}\sum_{t=2}^{T}\partial\big(\mathrm{vec}(\mathbf{S})^\top\big)\,\mathrm{vec}(\mathbf{Q}^{-1})/\partial\boldsymbol{q}^{*} &+ \partial\big(\frac{T-1}{2}\log|\mathbf{Q}^{-1}|\big)/\partial\boldsymbol{q}^{*}\\
\text{where } \mathbf{S} = \sum_{t=2}^{T}\big(&\widetilde{\mathbf{P}}_t - \widetilde{\mathbf{P}}_{t,t-1}\mathbf{B}^\top - \mathbf{B}\widetilde{\mathbf{P}}_{t-1,t} - \widetilde{\mathbf{x}}_t\mathbf{u}^\top - \mathbf{u}(\widetilde{\mathbf{x}}_t)^\top +\\
&\mathbf{B}\widetilde{\mathbf{P}}_{t-1}\mathbf{B}^\top + \mathbf{B}\widetilde{\mathbf{x}}_{t-1}\mathbf{u}^\top + \mathbf{u}(\widetilde{\mathbf{x}}_{t-1})^\top\mathbf{B}^\top + \mathbf{u}\mathbf{u}^\top\big)
\end{aligned}
\tag{91}
$$

Note, I have replaced $\log|\mathbf{Q}|$ with $-\log|\mathbf{Q}^{-1}|$. The determinant of a diagonal matrix is the product of its diagonal elements. Thus,

$$
\begin{aligned}
\partial\Psi/\partial\boldsymbol{q}^{*} = -\Big(&\frac{1}{2}\,\mathrm{vec}(\mathbf{S})^\top(\mathbf{f}^{*} + \mathbf{D}\boldsymbol{q}^{*})\\
&- \frac{T-1}{2}(\log(f_1^{*}) + \log(f_2^{*})...k\log(q_1^{*}) + l\log(q_2^{*})...)\Big)/\partial\boldsymbol{q}^{*}
\end{aligned}
\tag{92}
$$

where k is the number of times q_1 appears on the diagonal of $\mathbf{Q}$ and l is the number of times q_2 appears, etc. Taking the derivatives,

$$
\begin{aligned}
\partial\Psi/\partial\boldsymbol{q}^{*} &== \frac{1}{2}\mathbf{D}^\top\,\mathrm{vec}(\mathbf{S}) - \frac{T-1}{2}(\log(f_1^{*}) + ...k\log(q_1^{*}) + l\log(q_2^{*})...)/\partial\boldsymbol{q}^{*}\\
&= \frac{1}{2}\mathbf{D}^\top\,\mathrm{vec}(\mathbf{S}) - \frac{T-1}{2}\mathbf{D}^\top\mathbf{D}\boldsymbol{q}
\end{aligned}
\tag{93}
$$

$\mathbf{D}^\top\mathbf{D}$ is a $p\times p$ matrix with k, l, etc. along the diagonal and thus is invertable; as usual, p is the number of free elements in $\mathbf{Q}$. Set the left side to zero (a $1\times p$ matrix of zeros) and solve for $\boldsymbol{q}$. This gives us the update equation for $\mathbf{Q}$:

$$
\begin{aligned}
\boldsymbol{q}_{\text{new}} &= \frac{1}{T-1}(\mathbf{D}_q^\top\mathbf{D}_q)^{-1}\mathbf{D}_q^\top\,\mathrm{vec}(\mathbf{S})\\
\mathrm{vec}(\mathbf{Q})_{\text{new}} &= \mathbf{f}_q + \mathbf{D}_q\boldsymbol{q}_{\text{new}}
\end{aligned}
\tag{94}
$$

[8] Another, more common, way to do this is to use a "trace trick", $\mathrm{trace}(\mathbf{a}^\top\mathbf{A}\mathbf{b}) = \mathrm{trace}(\mathbf{A}\mathbf{b}\mathbf{a}^\top)$, to pull $\mathbf{Q}^{-1}$ out.

4.6.2 Special case: Q with one variance and one covariance

$$\mathbf{Q} = \begin{bmatrix} \alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha \end{bmatrix} \qquad \mathbf{Q}^{-1} = \begin{bmatrix} f(\alpha,\beta) & g(\alpha,\beta) & g(\alpha,\beta) & g(\alpha,\beta) \\ g(\alpha,\beta) & f(\alpha,\beta) & g(\alpha,\beta) & g(\alpha,\beta) \\ g(\alpha,\beta) & g(\alpha,\beta) & f(\alpha,\beta) & g(\alpha,\beta) \\ g(\alpha,\beta) & g(\alpha,\beta) & g(\alpha,\beta) & f(\alpha,\beta) \end{bmatrix}$$

This is a matrix with a single shared variance parameter on the diagonal and a single shared covariance on the off-diagonals. The derivation is the same as for the diagonal case, until the step involving the differentiation of $\log|\mathbf{Q}^{-1}|$:

$$\partial\Psi/\partial\boldsymbol{q}^{*} = \partial\Big(-\frac{1}{2}\sum_{t=2}^{T}\big(\operatorname{vec}(\mathbf{S})^{\top}\big)\operatorname{vec}(\mathbf{Q}^{-1}) + \frac{T-1}{2}\log|\mathbf{Q}^{-1}|\Big)/\partial\boldsymbol{q}^{*} \tag{95}$$

It does not make sense to take the partial derivative of $\log|\mathbf{Q}^{-1}|$ with respect to $\operatorname{vec}(\mathbf{Q}^{-1})$ because many elements of $\mathbf{Q}^{-1}$ are shared so it is not possible to fix one element while varying another. Instead, we can take the partial derivative of $\log|\mathbf{Q}^{-1}|$ with respect to $g(\alpha,\beta)$ which is $\sum_{\{i,j\}\in\text{set}_g}\partial\log|\mathbf{Q}^{-1}|/\partial\boldsymbol{q}^{*}{}_{i,j}$. Set g is those i,j values where $\boldsymbol{q}^{*} = g(\alpha,\beta)$. Because $g()$ and $f()$ are different functions of both α and β, we can hold one constant while taking the partial derivative with respect to the other (well, presuming there exists some combination of α and β that would allow that). But if we have fixed values on the off-diagonal, this would not be possible. In this case (see below), we cannot hold $g()$ constant while varying $f()$ because both are only functions of α:

$$\mathbf{Q} = \begin{bmatrix} \alpha & f & f & f \\ f & \alpha & f & f \\ f & f & \alpha & f \\ f & f & f & \alpha \end{bmatrix} \qquad \mathbf{Q}^{-1} = \begin{bmatrix} f(\alpha) & g(\alpha) & g(\alpha) & g(\alpha) \\ g(\alpha) & f(\alpha) & g(\alpha) & g(\alpha) \\ g(\alpha) & g(\alpha) & f(\alpha) & g(\alpha) \\ g(\alpha) & g(\alpha) & g(\alpha) & f(\alpha) \end{bmatrix}$$

Taking the partial derivative of $\log|\mathbf{Q}^{-1}|$ with respect to $\boldsymbol{q}^{*} = \left[\begin{smallmatrix} f(\alpha,\beta) \\ g(\alpha,\beta) \end{smallmatrix}\right]$, we arrive at the same equation as for the diagonal matrix:

$$\partial\Psi/\partial\boldsymbol{q}^{*} = \frac{1}{2}\mathbf{D}^{\top}\operatorname{vec}(\mathbf{S}) - \frac{T-1}{2}\mathbf{D}^{\top}\mathbf{D}\boldsymbol{q} \tag{96}$$

where again $\mathbf{D}^{\top}\mathbf{D}$ is a $p\times p$ diagonal matrix with the number of times $f(\alpha,\beta)$ appears in element $(1,1)$ and the number of times $g(\alpha,\beta)$ appears in element $(2,2)$ of $\mathbf{D}$; $p=2$ here since there are only 2 free parameters in $\mathbf{Q}$.

Setting to zero and solving for $\boldsymbol{q}^{*}$ leads to the exact same update equation as for the diagonal $\mathbf{Q}$, namely equation (94) in which $\mathbf{f}_q = 0$ since there are no fixed values.

4.6.3 Special case: a block-diagonal matrices with replicated blocks

Because these operations extend directly to block-diagonal matrices, all results for individual matrix types can be extended to a block-diagonal matrix with

those types:

$$\mathbf{Q} = \begin{bmatrix} \mathbb{B}_1 & 0 & 0 \\ 0 & \mathbb{B}_2 & 0 \\ 0 & 0 & \mathbb{B}_3 \end{bmatrix}$$

where $\mathbb{B}_i$ is any of the allowed matrix types, such as unconstrained, diagonal (with fixed or shared elements), or equal variance-covariance. Blocks can also be shared:

$$\mathbf{Q} = \begin{bmatrix} \mathbb{B}_1 & 0 & 0 \\ 0 & \mathbb{B}_2 & 0 \\ 0 & 0 & \mathbb{B}_2 \end{bmatrix}$$

but notice the entire block must be identical ($\mathbb{B}_2 \equiv \mathbb{B}_3$); one cannot simply share individual elements in different blocks. Either all the elements in two (or 3, or 4...) blocks are shared or none are shared.

This is ok:

$$\begin{bmatrix} c & d & d & 0 & 0 & 0 \\ d & c & d & 0 & 0 & 0 \\ d & d & c & 0 & 0 & 0 \\ 0 & 0 & 0 & c & d & d \\ 0 & 0 & 0 & d & c & d \\ 0 & 0 & 0 & d & d & c \end{bmatrix}$$

This is not ok:

$$\begin{bmatrix} c & d & d & 0 & 0 \\ d & c & d & 0 & 0 \\ d & d & c & 0 & 0 \\ 0 & 0 & 0 & c & d \\ 0 & 0 & 0 & d & c \end{bmatrix} \text{ nor } \begin{bmatrix} c & d & d & 0 & 0 & 0 \\ d & c & d & 0 & 0 & 0 \\ d & d & c & 0 & 0 & 0 \\ 0 & 0 & 0 & c & e & e \\ 0 & 0 & 0 & e & c & e \\ 0 & 0 & 0 & e & e & c \end{bmatrix}$$

The first is bad because the blocks are not identical; they need the same dimensions as well as the same values. The second is bad because again the blocks are not identical; all values must be the same.

4.6.4 Special case: a symmetric blocked matrix

The same derivation translates immediately to blocked symmetric $\mathbf{Q}$ matrices with the following form:

$$\mathbf{Q} = \begin{bmatrix} \mathbb{E}_1 & \mathbb{C}_{1,2} & \mathbb{C}_{1,3} \\ \mathbb{C}_{1,2} & \mathbb{E}_2 & \mathbb{C}_{2,3} \\ \mathbb{C}_{1,3} & \mathbb{C}_{2,3} & \mathbb{E}_3 \end{bmatrix}$$

where the $\mathbb{E}$ are as above matrices with one value on the diagonal and another on the off-diagonals (no zeros!). The $\mathbb{C}$ matrices have only one free value or are all zero. Some $\mathbb{C}$ matrices can be zero while are others are non-zero, but a individual $\mathbb{C}$ matrix cannot have a combination of free values and zero values;

they have to be one or the other. Also the whole matrix must stay block symmetric. Additionally, there can be shared $\mathbb{E}$ or $\mathbb{C}$ matrices but the whole matrix needs to stay block-symmetric. Here are the forms that $\mathbb{E}$ and $\mathbb{C}$ can take:

$$\mathbb{E}_i = \begin{bmatrix} \alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha \end{bmatrix} \qquad \mathbb{C}_i = \begin{bmatrix} \chi & \chi & \chi & \chi \\ \chi & \chi & \chi & \chi \\ \chi & \chi & \chi & \chi \\ \chi & \chi & \chi & \chi \end{bmatrix} \text{ or } \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The following are block-symmetric:

$$\begin{bmatrix} \mathbb{E}_1 & \mathbb{C}_{1,2} & \mathbb{C}_{1,3} \\ \mathbb{C}_{1,2} & \mathbb{E}_2 & \mathbb{C}_{2,3} \\ \mathbb{C}_{1,3} & \mathbb{C}_{2,3} & \mathbb{E}_3 \end{bmatrix} \text{ and } \begin{bmatrix} \mathbb{E} & \mathbb{C} & \mathbb{C} \\ \mathbb{C} & \mathbb{E} & \mathbb{C} \\ \mathbb{C} & \mathbb{C} & \mathbb{E} \end{bmatrix}$$

$$\text{and } \begin{bmatrix} \mathbb{E}_1 & \mathbb{C}_1 & \mathbb{C}_{1,2} \\ \mathbb{C}_1 & \mathbb{E}_1 & \mathbb{C}_{1,2} \\ \mathbb{C}_{1,2} & \mathbb{C}_{1,2} & \mathbb{E}_2 \end{bmatrix}$$

The following are NOT block-symmetric:

$$\begin{bmatrix} \mathbb{E}_1 & \mathbb{C}_{1,2} & 0 \\ \mathbb{C}_{1,2} & \mathbb{E}_2 & \mathbb{C}_{2,3} \\ 0 & \mathbb{C}_{2,3} & \mathbb{E}_3 \end{bmatrix} \text{ and } \begin{bmatrix} \mathbb{E}_1 & 0 & \mathbb{C}_1 \\ 0 & \mathbb{E}_1 & \mathbb{C}_2 \\ \mathbb{C}_1 & \mathbb{C}_2 & \mathbb{E}_2 \end{bmatrix} \text{ and } \begin{bmatrix} \mathbb{E}_1 & 0 & \mathbb{C}_{1,2} \\ 0 & \mathbb{E}_1 & \mathbb{C}_{1,2} \\ \mathbb{C}_{1,2} & \mathbb{C}_{1,2} & \mathbb{E}_2 \end{bmatrix}$$

$$\text{and } \begin{bmatrix} \mathbb{U}_1 & \mathbb{C}_{1,2} & \mathbb{C}_{1,3} \\ \mathbb{C}_{1,2} & \mathbb{E}_2 & \mathbb{C}_{2,3} \\ \mathbb{C}_{1,3} & \mathbb{C}_{2,3} & \mathbb{E}_3 \end{bmatrix} \text{ and } \begin{bmatrix} \mathbb{D}_1 & \mathbb{C}_{1,2} & \mathbb{C}_{1,3} \\ \mathbb{C}_{1,2} & \mathbb{E}_2 & \mathbb{C}_{2,3} \\ \mathbb{C}_{1,3} & \mathbb{C}_{2,3} & \mathbb{E}_3 \end{bmatrix}$$

In the first row, the matrices have fixed values (zeros) and free values (covariances) on the same off-diagonal row and column. That is not allowed. If there is a zero on a row or column, all other terms on the off-diagonal row and column must be also zero. In the second row, the matrix is not block-symmetric since the upper corner is an unconstrained block ($\mathbb{U}_1$) and diagonal block ($\mathbb{D}_1$) instead of a equal variance-covariance matrix ($\mathbb{E}$).

4.6.5 The general case: a block-diagonal matrix with general blocks

In it's most general form, $\mathbf{Q}$ is allowed to have a block-diagonal form where the blocks, here called $\mathbb{G}$ are any of the previous allowed cases. No shared values across $\mathbb{G}$'s; shared values are allowed within $\mathbb{G}$'s.

$$\mathbf{Q} = \begin{bmatrix} \mathbb{G}_1 & 0 & 0 \\ 0 & \mathbb{G}_2 & 0 \\ 0 & 0 & \mathbb{G}_3 \end{bmatrix}$$

The $\mathbb{G}$'s must be one of the special cases listed above: unconstrained, diagonal (with fixed or shared values), equal variance-covariance, block diagonal

(with shared or unshared blocks), and block-symmetric (with shared or unshared blocks). Fixed blocks are allowed, but then the covariances with the free blocks must be zero:

$$\mathbf{Q} = \begin{bmatrix} \mathbb{F} & 0 & 0 & 0 \\ 0 & \mathbb{G}_1 & 0 & 0 \\ 0 & 0 & \mathbb{G}_2 & 0 \\ 0 & 0 & 0 & \mathbb{G}_3 \end{bmatrix}$$

Fixed blocks must have only fixed values (zero is a fixed value) but the fixed values can be different from each other. The free blocks must have only free values (zero is not a free value).

4.7 The general R update equation

The $\mathbf{R}$ update equation for blocked symmetric matrices with optional independent fixed blocks is completely analogous to the $\mathbf{Q}$ equation. Thus if $\mathbf{R}$ has the form

$$\mathbf{R} = \begin{bmatrix} \mathbb{F} & 0 & 0 & 0 \\ 0 & \mathbb{G}_1 & 0 & 0 \\ 0 & 0 & \mathbb{G}_2 & 0 \\ 0 & 0 & 0 & \mathbb{G}_3 \end{bmatrix}$$

Again the $\mathbb{G}$'s must be one of the special cases listed above: unconstrained, diagonal (with fixed or shared values), equal variance-covariance, block diagonal (with shared or unshared blocks), and block-symmetric (with shared or unshared blocks). Fixed blocks are allowed, but then the covariances with the free blocks must be zero

The update equation is

$$\begin{aligned} \boldsymbol{\rho}_{\text{new}} = \frac{1}{T}(\mathbf{D}_r^\top \mathbf{D}_r)^{-1}\mathbf{D}_r^\top \operatorname{vec}\Bigg(\sum_{t=1}^{T} (\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})^\top \\ + \mathbf{Z}\widetilde{\mathbf{V}}_t\mathbf{Z}^\top \Bigg) \\ \operatorname{vec}(\mathbf{R})_{\text{new}} = \mathbf{f}_r + \mathbf{D}_r\boldsymbol{\rho}_{\text{new}} \end{aligned} \tag{97}$$

where $\mathbf{D}_r$ is the design matrix defined in the same way as $\mathbf{D}_q$ and $\mathbf{f}_r$ and $\boldsymbol{\rho}$ are column vectors of the fixed and free values defined in the usual way.

The update equation (97) is valid as long as there are no missing values that fall within free blocks. For example, if $y_{i,t}$ is missing, then $\mathbf{R}_{i,i}$ cannot be in a free block; it must be in a fixed block (an $\mathbb{F}$ block). If there are missing values that fall in free blocks, then the update equation above will not work unless $\mathbf{R}$ is diagonal. When $\mathbf{R}$ is diagonal (strictly diagonal, not block-diagonal) then the update equation (97) can be used with a missing value modification. If the i-th value of $\boldsymbol{y}$ is missing at time t, that is, element $y_{i,t}$, then the (i,i) value of $\mathbf{R}$ from the previous iteration of the EM algorithm, $\mathbf{R}_{(i,i),\text{old}}$, is used in place of the t-th (i,i) value in the matrix $(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})(\boldsymbol{y}_t - \mathbf{Z}\widetilde{\mathbf{x}}_t - \mathbf{a})^\top$ in the summation on the last line of equation (97).

5 Implementation comments

The EM algorithm is a hill-climbing algorithm and like all hill-climbing algorithms it can get stuck on local maxima. There are a number approaches to doing a pre-search of the initial conditions space, but a brute force random Monte Carol search appears to work well (Biernacki et al., 2003). It is slow, but normally sufficient. In our experience, Monte Carlo initial conditions searches become important as the fraction of missing data in the data set increases. Certainly an initial conditions search should be done before reporting final estimates for an analysis. However in our[9] studies on the distributional properties of parameter estimates, we rarely found it necessary to do an initial conditions search.

The EM algorithm will quickly home in on parameter estimates that are close to the maximum, but once the values are close, the EM algorithm can slow to a crawl. Some researchers start with an EM algorithm to get close to the maximum-likelihood parameters and then switch to a quasi-Newton method for the final search. In our ecological applications, parameter estimates that differ by less than 3 decimal places are for all practical purposes the same. Thus we have not used the quasi-Newton final search.

Shumway and Stoffer (2006) imply in their discussion of the EM algorithm that both $\boldsymbol{\xi}$ and $\mathbf{V}_1$ can be simultaneously estimated. Others have noted that the algorithm bogs down when one attempts this, and this has been our experience. Harvey (1989) discusses that there are only two allowable cases for the initial conditions: 1) fixed but unknown and 2) a initial condition set as a prior. In case 1, $\boldsymbol{\xi}$ is then estimated as a parameter and $\mathbf{V}_1$ is held fixed at 0. In case 2, neither $\boldsymbol{\xi}$ nor $\mathbf{V}_1$ are estimated. Rather they are specified, not estimated, as part of the model. In the Holmes and Ward (2010) paper, we use case 1.

For case 1, one cannot set $\mathbf{V}_1 = 0$ because $\boldsymbol{\xi}$ would never be able to leave the initial value – because you told it not to by setting its variance to zero. So, the algorithm won't work. If you try to circumvent this by setting $\mathbf{V}_1$ equal to some small, but not zero, value, the algorithm will work but it will be horribly slow. The solution, I found, is to set $\mathbf{V}_1$ to a large value, e.g. $\mathbf{V}_1 = \mathbf{I}_m$ where $\mathbf{I}_m$ is the $m \times m$ identity matrix. The final maximum-likelihood parameter values are unaffected by $\mathbf{V}_1$. Setting $\mathbf{V}_1 = \mathbf{I}_m$, lets the EM algorithm find the maximum-likelihood $\boldsymbol{\xi}$ value quickly. Once all the maximum-likelihood parameters are found via the EM algorithm, the algorithm reruns the Kalman filter[10] with the maximum-likelihood parameters and $\mathbf{V}_1 = 0$ to obtain the correct likelihood for case 1.

In some cases, the update equation for one parameter needs other parameters. Technically, the Kalman filter/smoother should be run between each parameter update, however following Ghahramani and Hinton (1996) our algorithm skips this step (unless the user sets `control$EMsafe=TRUE`) and each updated parameter is used for subsequent update equations.

[9] "Our" means work and papers by E. E. Holmes and E.J. Ward.

[10] Technically, the output from the Kalman filter is used in the 'innovations form of the likelihood' (eqn 4.67 in Shumway and Stoffer, 2006) to compute $\log L(\boldsymbol{y}_1^T|\hat{\Theta})$.

6 MARSS code package

R code for the Kalman filter, Kalman smoother, and EM algorithm is provided as a separate R package, MARSS. MARSS was developed by Elizabeth Holmes, Eric Ward and Kellie Wills and provides maximum-likelihood estimation and model-selection for both unconstrained and constrained MARSS models. The package contains a detailed manual which gives further information on the algorithms behind the likelihood computations, bootstrapping, confidence intervals, and model selection criteria.

References

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41(3-4):561–575.

Ghahramani, Z. and Hinton, G. E. (1996). Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge, UK.

Henderson, H. V. and Searle, S. R. (1979). Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *The Canadian Journal of Statistics*, 7(1):65–81.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*. John Wiley and Sons, Inc., Hoboken, NJ, 2nd edition.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural Computation*, 11:305–345.

Shumway, R. and Stoffer, D. (2006). *Time series analysis and its applications*. Springer-Science+Business Media, LLC, New York, New York, 2nd edition.

Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264.